KNOWLEDGE and its LIMITS

Timothy Williamson

OXFORD

Knowledge and its Limits

TIMOTHY WILLIAMSON

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Singapore Taipei Tokyo Toronto Warsaw

with associated companies in Berlin Ibadan

Published in the United States
by Oxford University Press Inc., New York

First published 2000

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Williamson, Timothy.
Knowledge and its limits / Timothy Williamson.
p. cm.
Includes bibliographical references and index.
1. Knowledge, Theory of. I. Title.
BD201 .W55 2000 121—dc21 00-038490
ISBN 0-19-825043-6 (alk. paper)

1 3 5 7 9 10 8 6 4 2

Typeset by Invisible Ink
Printed in Great Britain
on acid-free paper by
Biddles Ltd.,
Guildford & King's Lynn

PREFACE

If I had to summarize this book in two words, they would be: knowledge first. It takes the simple distinction between knowledge and ignorance as a starting point from which to explain other things, not as something itself to be explained. In that sense the book reverses the direction of explanation predominant in the history of epistemology.

Like many philosophers, I have long been impressed by the failure of attempts to find a correct analysis of the notion of knowledge in terms of supposedly more basic notions, such as belief, truth, and justification. One natural explanation of the failure is that knowledge has no such analysis. If so, I wondered, what follows? At first, I was tempted to draw the conclusion that the notion of knowledge did not matter very much, because we could use those other notions instead. Around 1986, however, I began to notice points at which philosophers had gone wrong through using combinations of those other notions when the notion of knowledge was what their purposes really called for. That raised the question: why did they not use the notion of knowledge when it was just what they needed? The first three chapters of this book explain but do not justify this neglect of the distinction between knowledge and ignorance. They do so by applying the lessons of recent philosophy of mind to epistemology and then using the result to enrich the philosophy of mind. That provides a theoretical context for work I had already been doing on knowledge and its limits, work in which the notion of knowledge figures as one of the main instruments of understanding. That work forms much of the basis for the final nine chapters. These chapters also sketch applications to the philosophy of language, the philosophy of science, and decision theory. The book suggests a way of doing epistemology in which the distinction between knowledge and ignorance is central and irreducible, and we can still aspire to systematicity and rigour.

This book draws together work done in many places. There are traces of my time at Trinity College Dublin and much more from that at University College Oxford, particularly from some periods of leave and partial teaching relief. The majority of the material is far more recent, since my move to the University of Edinburgh, again with valuable periods of leave and partial teaching relief. The hospitality of other institutions was also important: I did some of the work as a visiting professor at the Massachusetts Institute of Technology and Princeton University and as a visiting fellow of the Australian National University and the University of Canterbury.

Most of the ideas in the book have been tried out in discussion on many occasions, both informally and at graduate classes at Oxford, Edinburgh, Princeton, and Helsinki; talks at the University of Aberdeen, the Australian National University, the University of Belgrade, the University of Bristol, the Cambridge Moral Sciences Club, the University of Canterbury at Christchurch, Cornell University, the University of Delaware, the University of Edinburgh, the University of Glasgow, Keele University, La Trobe University, the University of Leeds, the Classical University of Lisbon, University College London, the Catholic University of Lublin, the Massachusetts Institute of Technology, Melbourne University, the Instituto de Investigaciones Filosóficas of the Universidad Nacional Autónoma de México, Monash University, the University of Michigan at Ann Arbor, the University of New Mexico, New York University, the University of North Carolina at Chapel Hill, Ohio State University, the University of Oslo, the University of Oxford, the University of St Andrews, the University of Sheffield, the University of Stirling, the University of Sussex, Waikato University, the University of Wollongong, and Yale University; workshops on epistemology at the Universities of London and Stirling; a conference in Glasgow on Achilles and the Tortoise; a conference on empiricism and a meeting of the Scots Philosophical Club, both at Edinburgh; the 1999 Rutgers conference on epistemology; a congress on analytic philosophy at the turn of the millennium at Santiago de Compostela. To anyone familiar with analytic philosophy, it hardly needs to be emphasized how much there is to be learned from such occasions. The reader must judge whether I have learned enough. Certainly some sections of the book emerged as answers to questions posed by members of one or more of those audiences. I thank those audiences collectively. In addition, individual thanks are due to many people: they include Michael Ayers, Michael Bacharach, Helen Beebee, Alexander Bird, Simon Blackburn, Bill Brewer, Justin Broackes, John Campbell, Peter Carruthers, Paul Castell, Bill Child, Tim Cleveland, Earl Conee, Jack Copeland, Neil Cooper, Paolo Crivelli, Jonathan Dancy, Keith DeRose, Harry Deutsch, Dorothy Edgington, Jim Edwards, Matti Eklund, Kit Fine, Graeme Forbes, Elizabeth Fricker, Richard Fumerton, Manuel Garcia Carpintero, Olav Gjelsvik, John Gibbons, Gilbert Harman, Pedro Hecht, James Higginbotham, Matthias Hild, Richard Holton, Lloyd Humberstone, Frank Jackson, Mark Johnston, Peter Klein, Jon Kvanvig, Igal Kvart, Rae Langton, Keith Lehrer, David Lewis, Peter Lipton, Michael Martin, Hugh Mellor, Peter Milne, Chad Mohler, Adam Morton, Peter Mott, Nicholas Nathan, John O'Leary-Hawthorne, Philip Percival, Philip Pettit, Stathis Psillos, Gideon Rosen, Mark Sainsbury, Nathan Salmon, Hyun Song Shin, Sydney Shoemaker, John Skorupski, Roy Sorensen, Ernest Sosa, Jason Stanley, Helen Steward, Scott Sturgeon, Richard Swinburne, Charles Travis, Peter Unger, Alan Weir, Ralph Wedgwood, Crispin Wright, and various anonymous referees. The lists are certainly both invidious and incomplete; I apologize to those whom I have undeservedly omitted, and hope that they will take some satisfaction from the improvements which they correctly guess themselves to have caused. Peter Momtchiloff has been helpful and supportive as my editor at Oxford University Press, and Angela Blackburn meticulous as my copy-editor. Elisabetta Williamson enabled me to spend an excessive proportion of my days writing the book. Alice and Conrad were Alice and Conrad.

At one stage I envisaged a collection of previously published papers, cluttered with additional footnotes and postscripts. Subsequently dissatisfied with that prospect, I reworked, expanded, and integrated the material. Some repetitions have been eliminated, terminology has been made uniform, and interconnections signalled. The original sources are listed below; the bibliography contains full details of the papers mentioned. I should have made few of these improvements had it not been for Mima Andjelković, who refused to believe that I had already done my best; she was right. She also caught many errors at proof stage.

The introduction is new.

Chapter 1 is based on parts of 'Is knowing a state of mind?', *Mind* 104 (1995), with extensive rewriting. There is significant new material in sections 1.1, 1.2, 1.3, and 1.5.

Most of Chapter 2 is based on parts of 'Is knowing a state of mind?', with some material in sections 2.1 and 2.2 from 'The broadness of the mental: some logical considerations', *Philosophical Perspectives* 12 (1998). There is extensive new material in section 2.3 and some in each of the other sections.

The majority of Chapter 3 is based on 'The broadness of the mental'. Section 3.2 contains significant new material, section 3.3 is largely new, and section 3.8 is wholly new.

Most of Chapter 4 is based on 'Cognitive homelessness', *The Journal of Philosophy* 93 (1996), with rewriting. Section 4.4 is new.

Of Chapter 5, sections 5.1 and 5.2 are based on parts of 'Inexact knowledge', *Mind* 101 (1992), with extensive reworking. The reworking differentiates those sections from sections 8.2 and 8.3 of my *Vagueness*, which were also based on 'Inexact knowledge'. Sections 5.4 and 5.5 are based on 'Margins for error: a reply', *Philosophical Quarterly* 50 (2000). Section 5.3 is new.

Chapter 6 is based on parts of 'Inexact knowledge'.

Most of Chapter 7 is new. Sections 7.4 and 7.5 overlap 'Skepticism, semantic externalism, and Keith's mom', *Southern Journal of Philosophy* 38 (2000).

Chapter 8 is mainly based on 'Scepticism and evidence', *Philosophy and Phenomenological Research* 60 (2000), with some additional material expanded from 'Is knowing a state of mind?' in section 8.2.

Chapter 9 is a revised version of 'Knowledge as evidence', *Mind* 106 (1997). There is significant new material in sections 9.2, 9.7, and 9.8.

Chapter 10 is based on 'Conditionalizing on knowledge', *British Journal for the Philosophy of Science* 49 (1998), except for section 10.5, which is based on part of 'Inexact knowledge'.

Chapter 11 is a revised version of 'Knowing and asserting', *Philosophical Review* 105 (1996), with some brief new passages.

Of Chapter 12, sections 12.1 and 12.2 draw on 'Verificationism and non-distributive knowledge', *Australasian Journal of Philosophy* 71 (1993), with amplifications and some reference to results in 'Two incomplete anti-realist modal epistemic logics', *Journal of Symbolic Logic* 55 (1990). Section 12.5 reworks 'On the paradox of knowability', *Mind* 96 (1987), and 'On knowledge of the unknowable', *Analysis* 47 (1987). Sections 12.3 and 12.4 are new.

Appendix 1 reprints the appendix to 'The broadness of the mental: some logical considerations'. Appendix 2 is a revised version of the appendix to 'Inexact knowledge'. Appendix 3 is new. Appendices 4 and 5 reprint the appendices to 'Conditionalizing on knowledge'. Appendix 6 is a revised version of the appendix to 'Verificationism and non-distributive knowledge'.

I thank the editors concerned (and Cornell University in the case of *Philosophical Review*) for permission to use this material.

T. W.

CONTENTS

INTRODUCTION

Everyone by nature desires to know.

Aristotle, *Metaphysics* A1 980a21 (modernized)

1 KNOWING AND ACTING

Knowledge and action are the central relations between mind and world. In action, world is adapted to mind. In knowledge, mind is adapted to world. When world is maladapted to mind, there is a residue of desire. When mind is maladapted to world, there is a residue of belief. Desire aspires to action; belief aspires to knowledge. The point of desire is action; the point of belief is knowledge.

Those slogans are not platitudes—unless platitudes can be generally contested. According to many philosophers, desire aspires only to satisfaction, and belief only to truth. Action is a systematic way to satisfied desire, and knowledge to true belief, but desires can also be satisfied and beliefs true by chance. There is satisfied desire without action and true belief without knowledge. Why ask for more? Satisfaction and truth already constitute the required match between mind and world, with the appropriate directions of fit. Of course, we sometimes desire to act; those desires are satisfied only if there is action. We sometimes believe ourselves to know; those beliefs are true only if there is knowledge. But such cases are special; our desires and beliefs frequently concern states of the world of which actions and beliefs are not themselves constituents.

Although desires can be satisfied as well by chance as by action, that is no reason to marginalize the category of action in the understanding of mind. The place of desire in the economy of mental life depends on its potential connection with action. Similarly, although beliefs can be true as well by chance as by knowledge, that is no reason to marginalize the category of knowledge in the understanding of mind. This book develops a conception on which the place of belief in the economy of mental life depends on its potential connection with knowledge.

The foregoing vague phrases will later be partially replaced by something more precise. But that is not the purpose of this introduction, which is painted with a broad brush. Its aim is to give the reader a rough overall picture in which the layout of the main parts is visible. Subsequent chapters fill in details in the parts. Even they will amount to nothing like a proof that the picture is correct. Epistemological theories are not usually susceptible of proof. This book shows how to understand cognitive phenomena on the basis of some simple but generally overlooked ideas. The reader will judge those ideas by their fruit.

2 UNANALYSABLE KNOWLEDGE

Contemporary accounts of mind tend to marginalize the category of knowledge, sometimes not mentioning it at all; they certainly make it less central than the category of action. As a reverse counterpart of the output from mind to world in action, they admit the input from world to mind in perception. The latter is a more restricted category than knowledge; it excludes the products of memory and conscious inference. Perception is the reverse counterpart of action if both are single episodes of causal interaction with the environment. But acting, in the sense of intentionally making something the case, includes far more complex and mediated adaptations of world to mind over extended periods. The reverse counterpart of action in that sense is knowledge. It includes far more complex and mediated adaptations of mind to world over extended periods than perception does.

On contemporary accounts of mind, the general category for states with the mind-to-world direction of fit is belief. The belief is true if it fits the world, false otherwise. Although true and false belief are the same mental state in different worlds, the place of belief in the economy of mental life depends on its potential connection with truth. Knowledge is merely a peculiar kind of true belief. Since Gettier showed that even justified true belief is insufficient for knowledge, epistemologists have expended vast efforts attempting to state exactly what kind of true belief knowledge is, but that programme is assigned no significance for the philosophy of mind. On such a view, knowledge is to be explained in terms of belief, and belief is what matters for the understanding of mind. The converse attempt to explain belief in terms of knowledge sounds eccentric and perverse. To summarize this orthodoxy: belief is conceptually prior to knowledge.

The orthodox claim is frequently taken for granted, rarely supported by argument. Why should we suppose that belief is conceptually prior to knowledge? One argument is that since knowledge entails belief but not vice versa, the entailment should be explained by the assumption that we conceptualize knowledge as the conjunction of belief with whatever must in fact be added to belief to yield knowledge—truth and other more elusive features. The conjuncts are conceptually prior to the conjunction. Given that knowledge entails belief, it is trivial that one knows *p* if and only if (1) one believes *p*; (2) *p* is true; and (3) if one believes *p* and *p* is true, then one knows *p*. But that equivalence is useless for establishing that belief is conceptually prior to knowledge, for it is circular: 'know' occurs in (3). The received idea is that we can conceptualize the factors whose conjunction with belief is necessary and sufficient for knowledge independently of knowledge; we can think of the former without already thinking of the latter, even implicitly. But the argument does not show that such independent conceptualization is possible, for a necessary but insufficient condition need not be a conjunct of a non-circular necessary and sufficient condition. Although being coloured is a necessary but insufficient condition for being red, we cannot state a necessary and sufficient condition for being red by conjoining being coloured with other properties specified without reference to red. Neither the equation 'Red = coloured + X' nor the equation 'Knowledge = true belief + X' need have a non-circular solution.

Thus belief can be a necessary but insufficient condition of knowledge even if we do not implicitly conceptualize knowledge as the conjunction of belief with that which must be added to belief to yield knowledge. Perhaps the inference from knowledge to belief derives from a conceptualization of belief in terms of knowledge rather than from a conceptualization of knowledge in terms of belief. If believing *p* is conceptualized as being in a state sufficiently like knowing *p* 'from the inside' in the relevant respects, then belief is necessary for knowledge, since knowing *p* is sufficiently like itself in every respect, even though knowledge is conceptually prior to belief. Indeed, the inference from knowledge to belief does not even require knowledge and belief to be conceptually ordered. We might master 'know' and 'believe' independently, from examples, and then realize on that basis that believing is necessary but insufficient for knowing, just as we might master the terms 'red' and 'scarlet' independently, from examples, and then realize on that basis that being red is necessary but insufficient for being scarlet. That belief is necessary but insufficent for knowledge does not show that belief is conceptually prior to knowledge. The orthodox claim would require a deeper defence.

Some epistemologists defend the conceptual priority of belief over knowledge by citing their favoured analyses of knowledge in terms of belief. Just what kind of conceptual priority such an analysis might support would depend on its status: for instance, on whether it was analytic or knowable a priori in some sense. But those issues become notional when, as usually happens, a counterexample is found to show that the proposed condition is not even necessary and sufficient for knowledge. Other analyses are circular rather than false; if someone insists that knowledge *is* justified true belief on an understanding of 'justified' strong enough to exclude Gettier cases but weak enough to include everyday empirical knowledge, the problem is likely to be that no standard of justification is supplied independent of knowledge itself. This book makes no attempt to survey even the most salient analyses of knowledge proposed in recent decades and the counterexamples to which they succumb; many other authors have already done that adequately. It will be assumed, not quite uncontroversially, that the upshot of that debate is that no currently available analysis of knowledge in terms of belief is adequate (not all parts of the book depend on that assumption). Consequently, the supposed conceptual priority of knowledge over belief is not to be defended by appeal to a particular analysis of knowledge in terms of belief.

A more cautious argument for the conceptual priority of belief over knowledge is that, even if all currently available analyses of knowledge in terms of belief are circular or fall to counterexamples, some of them are sufficiently good approximations to indicate strongly that a further refinement on similar lines will eventually succeed. But the possibility of approximating one concept with others is not good evidence that the former can be analysed in terms of the latter. For instance, to a very good approximation, x is a parent of y if and only if x is an ancestor of y and x is not an ancestor of an ancestor of y. The only counterexamples are *recherché* cases of incest: if a father incestuously begets a son on his daughter, the father is an ancestor of an ancestor of his son. But no more refined definition of parenthood in terms of ancestry alone avoids the problem. Since the father and the mother of his daughter are symmetrically related to the daughter and son in terms of ancestry but not in terms of parenthood, parenthood cannot be defined in terms of ancestry without extra conceptual resources. Moreover, the approximate definition of parenthood in terms of ancestry plays no significant role in our understanding of 'parent'. We can approximate a circle as closely as we like with sufficiently many sufficiently small triangles; it does not follow that we should think of the circle as made up out of triangles. The possibility of approximating knowledge in terms of belief and other concepts is not good evidence for the conceptual priority of belief over knowledge (section 1.3). Section 1.4 shows how one might characterize knowledge without reference to belief. Section 1.5 briefly discusses how one might characterize belief by reference to knowledge.

A chief aim of this book is to develop a rigorous way of doing epistemology in which knowledge is central, and not subordinate to belief. It enables us to abandon the attempt to state necessary and sufficient conditions for knowledge in terms of belief without abandoning epistemology itself. Indeed, by abandoning that fruitless search we can gain insight into epistemological problems, because we are freed to use the notion of knowledge as an instrument of understanding in ways that its subordination to belief would not permit (see, in particular, Chapter 9).

3 FACTIVE MENTAL STATES

The idea that belief is conceptually prior to knowledge has another source: the internalist conception of mind, and world external to mind, as two independent variables. Belief is simply a function of the mind variable. Truth is simply a function of the external world variable, at least when the given proposition is about the external world. For the internalist, knowledge is a function of the two variables, not of either one alone; whether one knows that it is raining does not depend solely on one's mental state, a state which is the same for those who perceive the rain and those who hallucinate it, but it also does not depend solely on the state of the weather, a state which is the same for those who believe the appearances and those who doubt them. The internalist therefore conceives knowledge as a complex hybrid crying out for analysis into its internal and external components, of which belief and truth respectively are the most salient. The analysis is expected on general metaphysical grounds.

Recent developments in the philosophy of mind have called the metaphysics of internalism into question by indicating ways in which the content of a mental state can constitutively depend on the environment. I believe that tigers growl; an exact physical replica of me lacks that belief if his contact has been not with tigers but with schmigers, beasts of a similar appearance belonging to a different species; his belief is that schmigers growl. Some internalists conclude that not even belief as attributed in ordinary language is simply a function of mind, and try in theory to isolate a core of purely mental states. Such attempts have not succeeded. Rather, we may conceive mind and external world as dependent variables, and reject the metaphysics that led us to expect analysis into purely internal and purely external components. On this view, belief as attributed in ordinary language is a genuine mental state constitutively dependent on the external world.

If the content of a mental state can depend on the external world, so can the attitude to that content. Knowledge is one such attitude. One's knowledge that it is raining depends on the weather; it does not follow that knowing that it is raining is not a mental state. The natural assumption is that sentences of the form 'S knows *p*' attribute mental states just as sentences of the forms 'S believes *p*' and 'S desires *p*' do. Chapters 1 and 2 defend such an externalist conception of knowing as a state of mind. In particular, section 2.3 refutes internalist attempts to isolate a non-factive state as the purely mental component of knowing.

What is at stake is much more than whether we apply the word 'mental' to knowing. If we could isolate a core of states which constituted 'pure mind' by being mental in some more thoroughgoing way than knowing is, then the term 'mental' might be extended to knowing as a mere courtesy title. On the conception defended here, there is no such core of mental states exclusive of knowing. If we want to illustrate the nature of mentality, knowing is as good an example as believing. The philosophy of mind cannot afford to neglect knowing, for that state is part of its core subject matter. For similar reasons, other truth-entailing attitudes such as perceiving and remembering that something is the case may also be classified as mental states. Knowing can be understood as the most general of such truth-entailing mental states (section 1.4).

Sceptics and their fellow-travellers characteristically suppose that the truth-values of one's beliefs can vary independently of those beliefs and of all one's other mental states: one's total mental state is exactly the same in a sufficiently radical sceptical scenario as it is in a common-sense scenario, yet most of one's beliefs about the external world are true in the common-sense scenario and false in the sceptical scenario. But if knowing is itself a mental state, that supposition is tantamount to the sceptical conclusion that in the common-sense scenario one's beliefs do not constitute knowledge, even though they are true. For, since false beliefs never constitute knowledge, one certainly does not know in the sceptical scenario; the supposition that one is in exactly the same mental state in the two scenarios therefore implies that one does not know in the common-sense scenario either, given that knowing makes a difference to one's total mental state (section 1.2). The anti-sceptic should not accept the supposition. Any mental life in the sceptical scenario is of a radically impoverished kind. Of course it does not *feel* impoverished 'from the inside', but that failure of self-knowledge is part of the impoverishment.

If action is the reverse counterpart of knowledge, and knowing is a mental state, should we expect acting to be a mental state too? If so, we might compare the sceptic's denial that we know about the external world to the denial that we act on the external world (perhaps made by those who believe that free will is both an illusion and a precondition of genuine action). But the analogy between knowledge and action is not perfect. Acting is by definition no state of any kind; it is dynamic, not static. Moreover, while knowing that the door is shut may be a mental state, shutting the door is surely not a mental action. Only actions such as inferring are naturally classified as mental. Similar asymmetries arise if we pursue the more restricted analogy between action and perception, for instance, between breaking the window and seeing that the window is broken. One starts seeing that the window is broken after the light rays reach one's retina; we do not make the apparently symmetric claim that one finishes breaking the window before the stone leaves one's hand. Why do we conceptualize the input and output sides so differently? The answer may lie in our tendency to individuate by origins. Effects depend on their causes in a way in which causes do not depend on their effects. Thus early stages in the process leading from a cause in the environment to a perceptual experience typically do not depend on the perceiver's involvement, whereas even late stages in the process leading from an intention to an effect on the environment do depend on the agent's involvement. Thus we naturally group both early and late stages of the output process into something attributable to the agent, while grouping only late stages of the input process into something attributable to the perceiver (this notion of grouping is intended to be neutral between different theories of the ontology of action). We treat early stages of the input process merely as preconditions for what we attribute to the perceiver. We extend this scheme to cases of knowledge and action with a more complex causal structure. Since late stages of the output process which occur without need of continued mental involvement are grouped into the action, we are reluctant to conceive it as mental. By contrast, since early stages of the input process are not grouped into the perception or knowledge, there is no corresponding block to conceiving it as mental.

Knowledge and action are related in another way. We expect genuine mental states to occur significantly in causal explanations of action, for otherwise postulating them looks redundant. Thus if knowing is a genuine mental state, it should occur significantly in such explanations. But many philosophers assume that attributions of knowledge in causal explanations of action can be replaced without explanatory loss by the corresponding attributions of belief. Section 2.4 and, in more depth, Chapter 3 undermine that assumption. Action typically involves complex interaction with the environment; one needs continual feedback to bring it to a successful conclusion. For example, writing a book involves reading during the process. Attributions of knowledge often explain the success of these interactions better than do the corresponding attributions of belief, even of true belief. One's belief in a proposition *p* is more robust to evidence if one knows *p* than if one merely believes *p* truly; one is less likely to lose belief in *p* in the course of interacting with the environment by discovering new evidence which lowers the probability of *p*. Thus one is more likely to complete an extended action that depends on a continuing belief in *p* if at the start one knows *p* rather than merely believes *p* truly. One is better placed to write a mathematical paper if one knows the truth of Goldbach's Conjecture than if one merely believes the conjecture and it is true. The point is not answered by an analysis of actions into series of basic actions, for the causal explanations even of basic actions often cite mental states at a temporal distance. One deliberates, forms an intention and then executes it later or abandons it in the light of new developments. The gap between deliberating and completing the action allows differences between knowledge and mere true belief in the basis of the deliberation to manifest themselves in action. If the causal explanation of the action cited only mental states immediately preceding the action, it would omit those on which the deliberation was based, and thereby miss the rationality of the action. These considerations can be generalized from attributions of knowledge to attributions of mental content involving reference to the environment; they all play distinctive roles in the causal explanation of temporally distant actions. Chapter 3 uses such considerations to argue that an externalist mental state normally cannot be decomposed as the conjunction of purely internal and purely external components.

4 KNOWLEDGE AS THE JUSTIFICATION OF BELIEF AND ASSERTION

The idea that belief is conceptually prior to knowledge easily leads to the idea that evidence and justification are conceptually prior to knowledge too. Although that is most vivid in the traditional definition of knowledge as justified true belief, Gettier's counterexamples to that definition did not remove the idea that the concept of justification or evidence would occur with the concept of belief in a more complex analysis of the concept of knowledge. Consequently, the concept of knowledge was assumed to be unavailable for use in an elucidation of the concept of justification or evidence, on pain of circularity. Once we cease to assume that belief is conceptually prior to knowledge, we can experiment with using the concept of knowledge to elucidate the concepts of justification and evidence.

Chapter 9 makes the experiment. It argues that one's total evidence is simply one's total knowledge. Thus a hypothesis is inconsistent with the evidence if and only if it is inconsistent with known truths; it is a good explanation of the evidence if and only if it is a good explanation of known truths. One's evidence justifies belief in the hypothesis if and only if one's knowledge justifies that belief. Knowledge figures in the account primarily as what justifies, not as what gets justified. Knowledge can justify a belief which is not itself knowledge, for the justification relation is not deductive. For example, I may be justified in believing that someone is a murderer by knowing that he emerged stealthily with a bloody knife from the room in which the body was subsequently discovered, even if he is in fact innocent and I therefore do not know that he is a murderer.

The equation of one's evidence with one's knowledge does not imply any particular theory of how a given body of propositional evidence justifies a given belief. Rather, it connects absolute and relative justification. A belief is justified relative to some other beliefs from which it has been derived in some appropriate way (perhaps by deduction), but it is not justified absolutely unless those other beliefs are justified absolutely. Where does the regress end? On the assumption that it ends at evidence, the equation of evidence with knowledge implies that one's belief is justified absolutely if and only if it is justified relative to one's knowledge. The regress of justification ends at knowledge.

The account might be thought to make all knowledge self-justifying in an absurdly trivial way: one's knowledge is justified absolutely if and only if it is justified relative to itself. This objection would be fair if the point of justification were to serve at its best as a condition for knowledge. But on the present account that is not the point of justification. Rather, justification is primarily a status which knowledge can confer on beliefs that look good in its light without themselves amounting to knowledge. Knowledge itself enjoys the status of justification only as a limiting case, just as, trivially, every shade of green counts as similar to a shade of green.

The objector might still point out that non-trivial questions appear to arise about the justification and evidence for much of our knowledge, especially that which is mediated by theory. We miss the specificity of these questions if we treat them merely as general questions about how we know. Nevertheless, they can be understood as non-trivial on the present approach. For even if one knows *p*, one can call that knowledge into question, provisionally treat *p* as though it did not belong to the body of one's knowledge, and then assess *p* relative to the rest of one's knowledge—one's independent evidence. Non-trivial issues of evidence and justification will then arise for *p*. This procedure is a good test of some kinds of supposed knowledge, especially those mediated by theory. In such cases, given the purported manner of knowing *p*, one knows *p* only if the rest of one's knowledge justifies *p*. But the test is not universal; it yields poor results if too much of one's knowledge is simultaneously called into question, for then one may easily have that knowledge even if removing it from the body of one's knowledge leaves too little to justify it. Some sceptics go wrong by applying the test in such cases.

The consideration of reduced bodies of evidence serves a different purpose. It enables us to isolate the contribution of specific pieces of evidence to the justification of specific hypotheses by comparing the status of those hypotheses relative to the total body of evidence with their status relative to the result of removing the piece of evidence in question from the total body of evidence This point is related to a form of the so-called problem of old evidence.

Chapter 10 develops these ideas in a more technical direction, by combining them with a theory of evidential probability in a modified objective Bayesian framework (some readers may prefer to skip this chapter). The evidential probability of a hypothesis for one is its probability conditional on one's total evidence; given the equation of one's evidence with one's knowledge, that is its probability conditional on one's total knowledge. Thus knowledge automatically receives evidential probability 1. Even so, knowledge is not treated as indefeasible evidence, for one can lose as well as gain knowledge. Thus the future evidential probability for me of my present knowledge may be less than 1. Together, Chapters 9 and 10 illustrate a way of doing epistemology on which knowledge is taken as the starting point, the unexplained explainer, yet some degree of rigour is maintained.

Chapter 11 extends the approach to the philosophy of language, with an account of the speech act of assertion. In a natural way we can regard assertion as the verbal counterpart of judgement and judgement as the occurrent form of belief. If one assumes that belief is conceptually prior to knowledge, one will therefore expect an account of assertion not to use the concept of knowledge. It might instead use the concepts of truth and justified belief, perhaps independently of each other. But if belief is not conceptually prior to knowledge, and knowledge is what justifies belief, then knowledge should play a key role in an account of assertion. On the proposal made in Chapter 11, the fundamental rule of assertion is that one should assert *p* only if one knows *p*. Although that knowledge rule might appear to be derivable from the truth rule that one should assert *p* only if *p* is true, by the consideration that in asserting *p* one does not know that one is conforming to the truth rule unless one is in fact conforming to the knowledge rule, the attempted derivation fails because it predicts the wrong epistemology for some examples. Even more incorrect predictions issue from an account of assertion based on the justification rule that one should assert *p* only if one is justified in believing *p*. The concept of knowledge is needed to capture our practice of assertion.

Given the combined conclusions of Chapters 9 and 11, the propositions which one is permitted to assert outright are exactly those which constitute one's evidence. More speculatively, we may project the account of assertion back onto its mental counterpart, judgement (or belief). What results is the rule that one should judge (or believe *p*) only if one knows *p*. That would make some sense of the claim that belief aims at knowledge. It also harmonizes with the account of evidence: to believe *p* without knowing *p* is to exceed one's evidence. Although we may have qualms about applying the notion of a rule to mental acts in addition to speech acts, the idea that belief is governed by a norm of knowledge is at least as intelligible as the idea that it is governed by a norm of truth.

5 THE MYTH OF EPISTEMIC TRANSPARENCY

An account has been sketched of knowledge as a mental state which constitutes the evidential standard for assertion and belief. The several components of the account face a common epistemological objection. It starts from the observation that one is not always in a position to know whether one knows something. If one knows *p*, it does not follow that one is in a position to know that one knows *p* (section 5.1); if one does not know *p*, it does not follow that one is in a position to know that one does not know *p* (section 8.2). In both cases, the conclusion fails to follow even if we add the extra premise that one is wondering whether one knows *p*; the problem is not confined to subjects who are unconscious, lack the concept of knowledge, or the like. In the simplest examples, one does not know and is not in a position to know that one does not know *p*. Sometimes *p* is false, so one does not know *p*, even though systematically misleading appearances place one in a state which feels just like knowing *p* 'from the inside', so one is not in a position to know that one does not know *p*. One falsely but justifiably believes oneself to know *p*. Examples in which one knows without being in a position to know that one knows will be discussed later. Why are these limitations on one's ability to know whether one knows supposed to threaten the foregoing account of knowledge?

Consider first the thesis that knowing is a mental state. We are often said to have special access to our own mental states, so that we can know without observation what mental states we are in. If S is a mental state only if one is always in a position to know whether one is in S (at least when one is in a position to wonder whether one is in S), then knowing is not a mental state.

A similar objection applies to the equation of evidence with knowledge. Rationality requires one to conform one's beliefs to one's evidence. Rationality cannot require one to do the impossible. But how can one conform one's beliefs to one's evidence unless one is in a position to know what it is? If one is always in a position to know what one's evidence is, then one's evidence is not one's knowledge.

Since one is not always in a position to know whether one knows *p*, one is not always in a position to know whether in asserting *p* one is complying with the rule 'Assert only what you know'. In particular, one may fail to meet the knowledge condition even though it feels 'from the inside' just as though one met the condition. A violation of the knowledge rule in such circumstances may look blameless. How can a speech act be governed by a rule that one can blamelessly violate? If one is always in a position to know whether one's assertions comply with the rule for assertion, then the rule is not 'Assert only what you know'.

If the first objection is sound, then every mental state has the property that one is in a position to know whether one is in it (the qualification 'whenever one is in a position to wonder whether one is in it' will henceforth often be left tacit). If the second and third objections are also sound, then mental states are qualified by their possession of that property to be both evidence and the standard for assertion, at least in respect of accessibility. Indeed, it is unclear how anything other than a mental state could be accessible in the required way (we may exclude trivial states which one is always or never in). For suppose that one is always in a position to know whether a condition C obtains. Consider an ordinary case α in which one might be and a sceptical counterpart α^* of α. In α^*, one is not in α but appears to oneself to be in α and for all one knows one is in α. If C obtains in α, then for all one knows in α^* one is in a situation in which C obtains; thus if C does not obtain in α^*, one is not in a position to know in α^* whether C obtains, contrary to hypothesis; therefore C obtains in α^*. By a parallel argument, if C does not obtain in α then C does not obtain in α^*. Thus C obtains in α if and only if C obtains in α^*. C is insensitive to the difference between ordinary cases and their sceptical counterparts. Mental states are the only obvious candidates for exhibiting such insensitivity.

We have also uncovered another temptation to scepticism, for if we combine the argument of the previous paragraph with the principle that one is always in a position to know what one's evidence is, the upshot is that one has exactly the same evidence in an ordinary case and its sceptical counterpart. How, then, can one know which case one is in?

The three objections assume that some non-trivial states meet the accessibility requirement; one is always in a position to know whether one is in them. Chapter 4 challenges that assumption. It provides a general form of argument, applicable to almost any condition, to undermine the claim that one is always in a position to know whether it obtains. More specifically, with rather trivial exceptions it undermines the claim that the condition is *luminous*, in the sense that whenever it obtains (and one is in a position to wonder whether it does), one is in a position to know that it obtains. The main idea behind the argument against luminosity is that our powers of discrimination are limited. If we are in a case α, and a case α' is close enough to α, then for all we know we are in α'. Thus what we are in a position to know in α is still true in α'. Consequently, a luminous condition obtains in α only if it also obtains in α', for it obtains in α only if we are in a position to know that it obtains in α. In other words, a luminous condition obtains in any case close enough to cases in which it obtains. What counts as close enough depends on our powers of discrimination. Since they are finite, a luminous condition spreads uncontrollably through conceptual space, overflowing all boundaries. It obtains everywhere or nowhere, at least where we are in a position to wonder whether it obtains. For almost any condition of interest, the cases in which it obtains are linked by a series of imperceptible gradations to cases in which it does not obtain, where at every step we are in a position to wonder whether it obtains. The condition is therefore not luminous. The full version of the argument cashes out those spatial metaphors in epistemic terms, to ensure that they do not import unwarranted presuppositions. In particular, the full version is formulated in a way which does not presuppose perfect sharpness in the boundary between the cases in which the condition obtains and the cases in which it does not. The upshot of the argument is that the gap between what is true and what we are in a position to know is not a special feature restricted to some problematic areas of discourse; it is normal throughout discourse.

Some may doubt the applicability of the argument against luminosity to mental states, on the grounds that it relies on a model of discrimination between independently constituted items, whereas one's mental states and one's judgements about them are held to be constitutively interdependent. But no such interdependence makes one's judgements about one's present mental states infallible. For instance, if my guru tells me that I shall feel intense pain at midnight and I am sufficiently gullible, I may judge at midnight that I am feeling intense pain; it does not follow that I am feeling intense pain. Of course, I may be in a position to know that I do not feel intense pain; my failure may be to actualize that potential. But the example still shows a gap between judgement and truth, even if smaller than elsewhere, which the argument can use as the thin end of a wedge against luminosity. The full version proceeds by analysis of the gradually varying degrees of confidence with which one judges, in a way applicable to judgements about mental states.

For virtually no mental state S is the condition that one is in S luminous. The condition that one is *not* in S is equally non-luminous. For example, one can love someone without being in a position to know that one loves them, and one can fail to love someone without being in a position to know that one fails to love them. One can want something without being in a position to know that one wants it, and one can fail to want something without being in a position to know that one fails to want it. Granted that knowing is a mental state, one should therefore not be surprised that one can fail to know something without being in a position to know that one fails to know it. Indeed, one can argue independently against luminosity for many mental states. They involve patterns of causal connections; sometimes one makes a judgement about one's present state which one is subsequently forced to retract, because one's intervening behaviour was in tension with the self-attributed pattern. One's judgements may be subject to systematic distortion. One's self-attributions of mental states are sometimes too unreliable to constitute knowledge. Mental states incompatible with one's self-image may be concealed from one. The difference between remembering an incident in one's early childhood and imagining it is a difference in mental state, but it is also one about which it is easy to be wrong.

None of this is to deny that in favourable cases one can know without observation whether one is in a given mental state. But knowledge meets that condition. You may know without observation whether you know that it rained two days ago, just as you may know without observation whether you believe that it rained two days ago. If you know that it rained two days ago, that knowledge (and belief) may result from past observations, but no *further* observations were needed to know that you know (and believe). Of course, subsequent observations indicating that it did not rain two days ago undermine the self-attribution of past knowledge that it rained two days ago without undermining the self-attribution of past belief that it rained two days ago. But if a judgement can be undermined by reasons of some kind, it does not follow that it was made on the basis of other reasons of the same kind. I can know without further observation that I know *p* even though observation can falsify a claim to know *p*.

Our extensive but not unlimited ability to know without further observation whether we know something is what enables us to use knowledge as evidence. It constitutes an extensive but not unlimited ability to know without further acquisition of evidence whether something is part of our present evidence. To complain that we are not always in a position to know whether we know something is to bankrupt the notion of evidence, for only luminous conditions meet that more stringent constraint, and luminous conditions are trivial. Although the constraint might drive us to suppose that one's evidence consists of appearances to oneself, the discrimination argument shows that not even the condition that things appear to one in a given way is luminous. For example, one may appear to oneself to be seeing a red patch even though one is not in a position to know that one appears to oneself to be seeing a red patch. Once the standard for the epistemic accessibility of evidence is set at an attainable level, knowledge meets the standard.

Chapter 8 traces the way in which excessive demands on the accessibility of evidence invite scepticism by diminishing evidence to an imaginary phenomenal substratum. If we presume to know too much about our evidence, we find ourselves knowing too little about the external world. The best argument for supposing that we have no more evidence in ordinary cases than in their sceptical counterparts trades on the false premise that the condition for being evidence is luminous. Since sceptics have not refuted the equation of evidence with knowledge, they are not entitled to assume that we have no more evidence in ordinary cases than in their sceptical counterparts, for on the view against which they are attempting to argue we do have more knowledge in ordinary cases than in their sceptical counterparts.

Since rationality requires one to conform one's beliefs to one's evidence, and one is not always in a position to know what one's evidence is, we need a conception of rationality on which we are not always in a position to know what it demands. Indeed, the anti-luminosity argument takes us more directly to the conclusion that one may be rationally required to do something even though one is not in a position to know that one is rationally required to do it. If we imagine that some candidate criterion of rationality is perfectly accessible, then we are always likely to prefer that criterion; but once we recognize that perfect accessibility is quite generally an unattainable ideal, we can learn to live with an imperfectly accessible criterion. We have nothing else to live with. Provided that one's evidence is more accessible than the truth-values of the hypotheses under investigation, the former can still serve as a useful guide to the latter. Real life is messy. Section 10.6 explores some unexpected implications of imperfect accessibility for decision theory. Imperfect accessibility has ethical implications too; we are not always in a position to know our duty.

For the same reason, one should expect not always to be in a position to know whether in asserting *p* one conforms to the rule of assertion. That the account of assertion based on the rule 'Assert only what you know' has that consequence is therefore no objection. An account based on the rule 'Assert only what you rationally believe' would have the same consequence. Our practice of assertion is workable because we often enough know whether we know something.

The imperfect accessibility of rationality casts light on the external individuation of mental content, mentioned earlier. For rationality has some relation to deductive logic, although the relation is not easy to spell out, and the external individuation of content makes the deductive validity of inferences imperfectly accessible. Whether the inference from 'It is hot here' and 'It is wet here' to 'It is hot and wet somewhere' is valid in a given context depends on whether the two occurrences of 'here' have the same content in that context. Someone who accepts the premises and rejects the conclusion avoids inconsistency only if the argument is invalid. If the content of 'here' is determined at least in part by the environment, one may not be in a position to know whether the inference is valid. Similarly, whether the inference from 'Everything changes' to 'Bourbaki changes' is valid depends on whether 'Bourbaki' has a content. If the content of 'Bourbaki' (if any) is determined at least in part by the environment, then one may not be in a position to know whether the inference is valid. These examples are only indicative; the argument from external individuation to imperfect accessibility is much less straightforward if the externally individuated content of a term is not identified with its referent. Nevertheless, if we accept on independent grounds that one is not always in a position to know what rationality demands, we should not object to an account that individuates content externally just on the grounds that it makes validity imperfectly accessible.

Chapter 5 articulates the constraints on knowledge implicit in the argument against luminosity. Where one has only a limited capacity to discriminate between cases in which *p* is true and cases in which *p* is false, knowledge requires a margin for error: cases in which one is in a position to know *p* must not be too close to cases in which *p* is false, otherwise one's belief in *p* in the former cases would lack a sufficiently reliable basis to constitute knowledge. The kind and degree of closeness in question depend on the specific limitations of one's powers of discrimination in that context. Thus the area of conceptual space in which one is in a position to know *p* is separated from the surrounding area in which *p* is false by a border zone in which *p* is true but one is not in a position to know *p*. The implications of the model are explored for iterated knowledge. In particular, one has only a limited capacity to discriminate between cases in which one knows *p* and cases in which one does not know *p*, so one can know *p* without being in a position to know that one knows *p*. Further iterations of knowledge are even harder to achieve. Naturally, one has an even more limited capacity to discriminate between cases in which others know *p* and cases in which they do not know *p*, so it is even harder to achieve iterations of shared knowledge ('We all know that we all know that we all know . . . *p*'), and a fortiori to achieve the infinitely many levels of iteration required for common knowledge. Chapter 6 uses this difficulty to account for the Paradox of the Unexpected Examination and some paradoxical arguments in game theory which assume that it is common knowledge amongst the players that they are all rational.

Because we need margins for error, it is implicit in our practice of assertion that truth outruns warranted assertion. We are warranted in asserting *p* only if we know *p*; we know *p* only if *p* is true in nearby cases. To interpret our assertions as warranted, we must interpret their content as true in some cases in which we are not warranted in asserting it. Our ignorance is a precondition of our knowledge. Contrary to anti-realist theories, the gap between assertibility and truth is built into even the simplest kinds of assertion.

Chapter 7 exploits margins for error in another direction. Sometimes one knows *p* by believing *p* and leaving a large margin for error even though, if *p* were false, one would still believe *p*. For one's judgement concerning *p* may be almost completely accurate but subject to a very slight distortion: when *p* is false but very close to being true, one falsely believes *p*. Since one's belief in *p* leaves a wide margin for error, it would have been false only if things had been very different; it may well be that if *p* had been false, it would still have been very close to being true. For example, if someone is very much less than two metres tall, I may know by sight that she is less than two metres tall, even though, if she were not less than two metres tall, she would be only very slightly more than two metres tall, and I would falsely judge her by sight to be less than two metres tall. Such cases falsify accounts of knowledge on which a necessary condition for knowing *p* is that if *p* were false one would not believe *p*. Sophisticated modifications of such accounts are refuted by cases in which the distortion in judgement is slight but ubiquitous. This result bears on some sceptical arguments, for we can take *p* to be the proposition that I am not in a sceptical scenario α. If *p* were false I would be in α and (by construction of α) would believe that I was not in α, but that does not justify the sceptical claim 'I do not know that I am not in α', for the counterfactual condition is not necessary for knowledge. Although it can still be insisted that a necessary condition for knowing *p* is that if *p* were false one would not believe *p* on the very same evidence, that counterfactual does not justify the sceptical claim, for the sceptic has not shown that one would have the very same evidence in the sceptical scenario. Given the account in chapter 9, in a sceptical scenario one's evidence is so radically impoverished that one is not in a position to know that it is impoverished at all.

Margins for error constitute a kind of epistemic friction. For some purposes it is useful to idealize them away in thought experiments, but a world in which there were no margins for error would be as different from our world as would a world in which there was no friction. We have no more reason to postulate that there really is a kind of knowledge of temporal matters without margins for error than we have to postulate that somewhere in space there really is a frictionless plane.

6 UNKNOWABLE TRUTHS

When knowing *p* requires a margin for error, the cases in which *p* is known are separated from the cases in which *p* is false by a buffer zone, a protective belt of cases in which *p* is true but unknown. That belt has the peculiarity that one cannot know that one is in it. For to know that would be to know that *p* is true and unknown; but knowing that involves knowing that *p* is true (since knowing a conjunction involves knowing its conjuncts); then *p* is not unknown, so it is not true that *p* is true and unknown, so it is not known that *p* is true and unknown (since only truths are known). Thus it is impossible to know that *p* is true but unknown. When *p* is in the protective belt, that is an unknowable truth.

The limits on knowledge in question are of a stronger kind than anything established by the anti-luminosity argument of Chapter 4. A proposition requires a margin for error precisely so that it can be known; the point of the anti-luminosity argument is just that the cases in which *p* is available to be known do not exhaust the cases in which *p* is true. By contrast, the point about the conjunctive proposition that *p* is true and unknown is that, in virtue of its structure, it is not available to be known in any case whatsoever. The argument for this conclusion was first published by Fitch in 1963. Contrapositively, he showed that all truths are knowable only if all truths are known. This is sometimes known as the Paradox of Knowability, although why it should be thought to constitute a paradox is unclear. It is the topic of Chapter 12.

Attempts have been made to take the sting out of Fitch's argument. Although you cannot know today the conjunction that *p* is true and you do not know *p* today, you can know the conjunction that *p* is true and you did not know *p* yesterday, and you can know the conjunction that *p* is true and I do not know *p* today. Thus one might distinguish a context in which Fitch's conjunction is true from a context in which its truth in the former context is known. The immediate response is to generalize the second conjunct, as in '*p* is true and no one ever knows *p*'. Fitch's argument shows that no one can ever know the conjunction that *p* and no one ever knows *p*. Many truths are never known by anyone. For example, either it is true that I had an even number of books in my office exactly a year ago or it is true that I had an odd number of books in my room exactly a year ago; no one will ever know which, because they were not counted at the time and it is now too late to find out. Nevertheless, one might try to distinguish a possible world *w* in which the conjunction (that *p* is true and no one ever knows *p*) is true from a possible world *w** in which the truth of the conjunction in *w* is known. The trouble with this move is that it promises only trivial knowledge. We specify merely possible worlds by description; in *w** we can describe a world as one in which *p* is true, and thereby know that *p* is true in such a world, but that is hardly a notable achievement. Section 12.5 argues that this knowledge of other possible worlds does not significantly relax the limits on knowability that Fitch's argument identifies.

Section 12.2 discusses a more direct challenge. Fitch's argument uses the distribution principle that knowledge of a conjunction implies knowledge of its conjuncts. Although the principle sounds compelling, a few accounts of knowledge are inconsistent with it. Probably that indicates something wrong with those accounts. As a precaution, ways are explored of modifying Fitch's argument to avoid relying on the distribution principle.

Limits on knowledge have counterparts in limits on action. On at least one interpretation, the relation between an agent and a proposition of *making true* shares the formal features of knowing needed for Fitch's argument: it distributes over conjunction (one makes a conjunction true only if one makes its conjuncts true) and is factive (if one makes something true then it is true). Thus no one can ever make this conjunction true: *p* and no one ever makes *p* true. For if one makes the conjunction true, one makes the first conjunct true, so the second conjunct is false, so the conjunction is false, so one did not make it true after all. If some truths were not made true by anyone, then some truths could not have been made true by anyone. Much of the world is outside both our control and our ken. We should find the limits on our knowledge scarcely more surprising than the limits on our action. Although knowledge and action are central to mind, mind is not central to world.

1
A State of Mind

1.1 FACTIVE ATTITUDES

Knowing is a state of mind. That claim is central to the account of knowledge developed in this book. But what does it mean?

A state of a mind is a mental state of a subject. Paradigmatic mental states include love, hate, pleasure, and pain. Moreover, they include attitudes to propositions: believing that something is so, conceiving that it is so, hoping or fearing that it is so, wondering whether it is so, intending or desiring it to be so. One can also know that something is so. This book concerns such propositional knowledge. If *p* is a proposition, we will understand knowing *p* not as merely being acquainted with *p* but as knowing that something is so, something that is so if and only if *p* is true. For example, if *p* is the proposition that it is cold, then one is acquainted with *p* in merely wondering whether it is cold; to know *p* is to know that it is cold. Knowing in that sense is a *factive* attitude; one knows *p* only if *p* is true, although one can be acquainted with the proposition *p* even if it is false. Other factive attitudes include perceiving that something is so, remembering that it is so, and regretting that is so. If attitudes are relations of subjects to propositions, then the claim is that knowing itself is a mental relation such that, for every proposition *p*, having that relation to *p* is a mental state. Thus for some mental state S, being in S is necessary and sufficient for knowing *p*. We abbreviate that claim by saying that knowing is a mental state.

We may assume initially that knowing *p* entails believing *p*; section 1.5 considers that assumption in more depth. Someone might expect knowing to be a state of mind simply on the grounds that knowing *p* involves the paradigmatic mental state of believing *p*. If those grounds were adequate, the claim that knowing is a state of mind would be banal. However, those grounds imply only that there is a mental state being in which is *necessary* for knowing *p*. By contrast, the claim that knowing is a state of mind is to be understood as the claim that there is a mental state being in which is necessary *and sufficient* for knowing *p*. In short, knowing is *merely* a state of mind. This claim may be unexpected. On the standard view, believing is merely a state of mind but knowing is not, because it is factive: truth is a non-mental component of knowing.

Our initial presumption should be that knowing is a mental state. Prior to philosophical theory-building, we learn the concept of the mental by examples. Our paradigms should include propositional attitudes such as believing and desiring, if our conception of the mental is not to be radically impoverished. But factive attitudes have so many similarities to the non-factive attitudes that we should expect them to constitute mental states too; we expect a concept to apply to whatever sufficiently resembles its paradigms. It would be strange if there were a mental state of fearing but no mental state of regretting, or a mental state of imagining but no mental state of remembering. Indeed, it is not clear that there are any pretheoretic grounds for omitting factive attitudes from the list of *paradigmatic* mental states. That the mental includes knowing and other factive attitudes is built into the natural understanding of the procedure by which the concept of the mental is acquired. Of course, that does not exclude the subsequent discovery of theoretical reasons for drawing the line between the mental and the non-mental somewhere else. But the theory behind those reasons had better be a good one.

This chapter and the next eliminate some putative differences between knowing and non-factive attitudes that might be thought to disqualify knowing as a mental state. The supposed disqualifications concern constitutive dependence on the environment, first-person accessibility, and causal efficacy. In each case, the differences dissolve on inspection. Naturally, this form of argument cannot provide conclusive proof. We survey the current candidates and find them wanting. We can still wonder whether our list of potential differences is complete. But without good theoretical reasons to demote knowing from its pretheoretical status as a central case of a mental state, demotion is surrender to mere special pleading. Indeed, conceptions on which knowing is the wrong kind of state to count as mental are objectionable on independent grounds. We can best understand knowing by classifying it with other mental phenomena.

In this chapter, section 1.2 orients the claim that knowing is a mental state with respect to some traditional issues about scepticism and self-knowledge. Section 1.3 explains an incompatibility between the view of knowing as a factive mental state and standard analyses of the concept *knows* as a conjunction of the concepts *believes* and *true* (predicated of the proposition) and of other concepts; it blames the analyses. Section 1.4 presents a modest positive account of the concept *knows*, distinguishes it from analyses of the traditional kind, and indicates the possibility of understanding epistemology in terms of the metaphysics of states. Section 1.5 discusses the relation between knowing and believing, and explores some implications for so-called disjunctive accounts of mental states.[1]

1.2 MENTAL STATES, FIRST-PERSON ACCESSIBILITY, AND SCEPTICISM

The conception of knowing as a mental state can look like a confusion between objective and subjective certainty. Someone might even diagnose that conception as Descartes' central mistake. Did he not seek a mental state sufficient for knowing *p*? Was not clearly and distinctly conceiving *p* his candidate? And does not the failure of his epistemological programme manifest the impossibility of a mental state of the required kind?

On the view to be developed here, if Descartes sought a mental state sufficient for knowing, his mistake lay elsewhere: perhaps in the view (if he held it) that one must always be in a position to know what mental state one is in. H. A. Prichard, who also took knowing to be a mental state, held that one is always in a position to know whether one knows or merely believes (Prichard 1950: 86). Few would now claim such powers of discrimination. Indeed, one cause of denials that knowing is a mental state may be the assumption that one must always be in a position to know whether one is in a given mental state.

One is surely not always in a position to know whether one knows *p* (for almost any proposition *p*), however alert and conceptually sophisticated one is. The point is most vivid when the subject believes *p* falsely. Consider, for example, the situation of a generally well-informed citizen N.N. who has not yet heard the news from the theatre where Lincoln has just been assassinated. Since Lincoln is dead, he is no longer President, so N.N. no longer knows that Lincoln is President (knowing is factive). However, N.N. is in no position to know that anything is amiss. He continues reasonably to believe that Lincoln is President; moreover, this seems to him to be just another item of general knowledge. N.N. continues reasonably to believe that he knows that Lincoln is President. Although N.N. does not know that Lincoln is President, he is in no position to know that he does not know that Lincoln is President (see also Hintikka 1962: 106 and section 8.2).

[1] McDowell 1995 and Gibbons 1998 defend closely related conceptions of knowing as a mental state. See also Guttenplan 1994 and Peacocke 1999: 52–5.

The argument as stated assumes that no a priori reasoning demonstrates that it is impossible to have knowledge about the external world, for such reasoning would make it unreasonable for N.N. to believe that he knows that Lincoln is President. Of course, if all knowledge is impossible then, for any proposition *p* whatsoever, one does not know *p* and is not in a position to know that one fails to know *p*; one is never in a position to know whether one knows *p*. A sceptic about the external world who is not a sceptic about everything might attempt to maintain that, for any informative proposition *p* about the external world, one is in a position to know that one does not know *p*. Let us assume for the time being that such a sceptic is wrong. Chapter 8 will reconsider scepticism.

We can also construct cases in which one knows *p* without being in a position to know that one knows *p* (see Chapter 5). They involve more delicate issues. It is enough for present purposes that one can fail to know *p* without being in a position to know that one fails to know *p*.

Let transparency be the thesis that for every mental state S, whenever one is suitably alert and conceptually sophisticated, one is in a position to know whether one is in S. Given transparency, knowing *p* is not a mental state, for almost any proposition *p*.

Transparency is false, however, and demonstrably so by reference to uncontentiously paradigmatic mental states. For example, one is sometimes in no position to know whether one is in the mental state of hoping *p*. I believe that I do not hope for a particular result to a match; I am conscious of nothing but indifference; then my disappointment at one outcome reveals my hope for another. When I had that hope, I was in no position to know that I had it. Indeed, it is hard to find a non-trivial mental state for which transparency holds. It fails for the state of believing *p*, for the difference between believing *p* and merely fancying *p* depends in part on one's dispositions to practical reasoning and action manifested only in counterfactual circumstances, and one is not always in a position to know what those dispositions are. Transparency is even doubtful for the state of being in pain; with too much self-pity one may mistake an itch for a pain, with too little one may mistake a pain for an itch. A form of argument will be developed in Chapter 4 to show that *no* non-trivial mental state satisfies transparency. But even if transparency does hold for a few mental states, it clearly fails for others; the premise of the argument from transparency to the denial that knowing *p* is a mental state is false. Given that knowing *p* is a mental state, we will not expect knowing whether one is in it to be always easy.

It does not follow that there is no asymmetry at all between knowledge of one's own mental states and knowledge of the mental states of others. Perhaps failures of transparency could not be the normal case, although that claim would require extensive argument. A more plausible claim is that we have some non-observational knowledge of our own mental states and not of the mental states of others. But then the same may be said of knowing: we have some non-observational knowledge of our own knowledge and ignorance and not of the knowledge and ignorance of others. Any genuine requirement of privileged access on mental states is met by the state of knowing *p*. Knowing is characteristically open to first-person present-tense access; like other mental states, it is not perfectly open.

Some may object that knowing whether one knows *p* requires evaluating reasons for and against *p* in a way in which knowing whether one believes *p* does not. They distinguish knowing whether one currently believes *p* from deciding whether to continue believing *p*. Suppose for a moment that they are correct in taking knowing whether one believes *p* not to require one to evaluate reasons for and against *p*. Still, even on their view there is also the mental state of *rationally* believing *p*, on some appropriate concept of rationality. Knowing whether one rationally believes *p* does require one to evaluate reasons for and against *p*. Thus the need for such evaluation in order to know whether one knows *p* does not show that knowing *p* is not a mental state.

Could it be replied that knowing and rationally believing are not mental states in the way that believing is, because 'know' and 'rational' are normative terms? Belief attributions have a normative element too, for to have any mental attitude to a content one must in some sense grasp that content, and therefore have some minimal ability to deal rationally with it; the reply itself classifies 'rational' as a normative term. In any sense in which 'know' and 'rational' are normative terms, ascriptions of mental states can be normative.

A different objection is that one's belief about whether one knows *p* is defeasible by new information in a way in which one's belief about whether one believes *p* is not. For example, the new information might show that *p* is false. But is one's belief about whether one believes *p* really indefeasible by new information? Someone might believe that he believes that the world will end next year, because he has joined a religious sect in which there is strong pressure to believe that the world will end next year, but his unwillingness to cash in his pension may suggest that he does not really believe that the world will end next year. When he reflects on his unwillingness to cash in his pension, he may come to that conclusion himself. But even if we forget such examples and suppose that one's belief about whether one believes *p* is not defeasible by further evidence, we must still acknowledge mental states such as being alert or thinking clearly about a problem. One's belief about whether one is alert or thinking clearly about a problem is defeasible by new information, for example about what drugs had been slipped into one's drink. Thus the defeasibility of beliefs about whether one knows *p* does not show that knowing *p* is not a mental state.

Once we consider the full variety of acknowledged mental states, it is clear that any general requirements of privileged access on mental states are very mild. Knowing satisfies those mild requirements.

The failure of transparency helps to clarify the relation between the thesis that knowing is a mental state and a traditional pattern of sceptical argument. The sceptic argues that a subject with a true belief could have been in exactly the same mental state (that is, in the same total set of mental states) even if the belief had been false. He concludes that, since the belief fails to constitute knowledge in the latter case, it fails equally to do so in the former. The sceptical argument assumes something like this: if one's mental state is exactly the same in two situations, then one's knowledge is also the same. On the account to be developed here, that assumption is correct, although not quite in the way that the sceptic imagines.

The sceptic supposes that a difference in knowledge would require some *prior* difference in mental state, which the subject could detect. On the present account, a difference in knowledge would *constitute* a difference in mental state. This difference need not be detectable by the subject who lacks knowledge. Thus the sceptic's assumption is correct for reasons that undermine his argument. He claims to have constructed a case in which the belief is false although the mental state is exactly the same. But the most that he has really shown about the case is that the belief is false and one's situation is not discriminably different. He has not shown that one cannot be in different mental states in indiscriminable situations. Indeed, since we are sometimes in no position to know whether we are in a given mental state, as argued above, surely one can be in different mental states in situations between which one cannot discriminate (see Chapter 8 and McDowell 1982).

If knowing is a mental state, then the sceptical argument is not compelling. Indeed, such a view of knowledge need only be defensible for the sceptical argument not to be compelling. Thus *one* route into scepticism is blocked. It is not the purpose of this chapter to argue that all are. Chapter 8 will consider sceptical reasoning more carefully.

If someone has already taken the route into scepticism offered by that fallacious argument, before it was blocked, and has become genuinely undecided, at least in principle, as to whether she is in a sceptical scenario, then the blocking of the route now comes too late to rescue her.

Nothing said here should convince someone who has given up ordinary beliefs that they did in fact constitute knowledge, for nothing said here should convince her that they are true. The trick is never to give them up. This is the usual case with philosophical treatments of scepticism: they are better at prevention than at cure. If a refutation of scepticism is supposed to reason one out of the hole, then scepticism is irrefutable. The most to be hoped for is something which will prevent the sceptic (who may be oneself) from reasoning one into the hole in the first place.

The purpose of these remarks has been to give a feel for the view that knowing is a state of mind. The content of the view must now be examined more explicitly. The notion of a mental state will not be formally defined, for that would require a formal definition of the mental. Rather, reflection on the intuitive notion of a mental state will help to clarify its workings. Section 1.4 will provide a less informal account.

1.3 KNOWLEDGE AND ANALYSIS

To call knowing a mental state is to assimilate it, in a certain respect, to paradigmatic mental states such as believing, desiring, and being in pain. It is also to contrast it with various non-examples of mental states. Perhaps the most revealing contrast is between knowing and believing truly.

Believing *p* truly is not a mental state, at least, not when *p* is an ordinary contingent proposition about the external environment. Intuitively, for example, there is no mental state being in which is necessary and sufficient for believing truly that it is raining (that is, for believing while it is raining that it is raining), just as there is no mental state being in which is necessary and sufficient for believing while Rome burns that it is raining. There is a mental state of believing that it is raining, and there is—on the present account—a mental state of knowing that it is raining, but there is no intermediate mental state of believing truly that it is raining. Let S_1 be knowing that it is raining, S_2 be believing truly that it is raining, and S_3 be believing that it is raining. Then, we may assume, necessarily, everything that is in S_1 is in S_2; necessarily, everything that is in S_2 is in S_3. Nevertheless, on the present account, although S_1 and S_3 are mental states, S_2 is not a mental state.

That something sandwiched between two mental states need not itself be a mental state is not as paradoxical as it may sound. Consider an analogy: the notion of a geometrical property. For these purposes, we can understand geometrical properties to be properties possessed by particulars in physical space. Let π_1 be the property of being an equilateral triangle, π_2 the property of being a triangle whose sides are indiscriminable in length to the naked human eye, and π_3 the property of being a triangle. Necessarily, everything that has π_1 has π_2, because lines of the same length cannot be discriminated in length; necessarily, everything that has π_2 has π_3. Nevertheless, although π_1 and π_3 are geometrical properties, π_2 is not a geometrical property, because it varies with variations in human eyesight. Something sandwiched between two geometrical properties need not itself be a geometrical property. Similarly, there is no structural reason why something sandwiched between two mental states should itself be a mental state.

The point is general. If S is a mental state and C a non-mental condition, there need be no mental state S* such that, necessarily, one is in S* if and only if one is in S and C obtains. The non-existence of such an S* is quite consistent with the existence of a mental state S** such that, necessarily, one is in S** only if (but not: if) one is in S and C is met. A mental state can guarantee that conjunction only by guaranteeing more than that conjunction.

If the denial that believing truly is a mental state does not immediately convince, think of it this way. Even if believing truly is a mental state in some liberal sense of the latter term, there is also a more restrictive but still reasonable sense in which believing truly is not a mental state but the combination of a mental state with a non-mental condition. The present claim is that knowing is a mental state in *every* reasonable sense of that term: there is no more restrictive but still reasonable sense of 'mental' in which knowing can be factored, like believing truly, into a combination of mental states with non-mental conditions. A sense of 'mental' is reasonable if it is sufficiently close to an ordinary sense of the word in important respects. Although the present claim is therefore vague, it is at least clear enough to be disputed.

Strictly speaking, we must distinguish a conceptual and a metaphysical contrast. The conceptual contrast is that the concept *knows* is a mental concept while the concept *believes truly* is not a mental concept. The metaphysical contrast is that knowing is a mental state while believing truly is not a mental state.

The concept *mental state* can at least roughly be defined in terms of the concept *mental concept of a state*: a state is mental if and only if there could be a mental concept of that state. This definition does not in principle exclude the possibility of a non-mental concept of a mental state, for different concepts can be of the same state. We may reasonably assume that states S_1 and S_2 are identical if and only if necessarily everything is in S_1 if and only if it is in S_2. In a given context, distinct concepts may be necessarily coextensive. For example, since gold is necessarily the element with atomic number 79, the state of having a tooth made of gold is the state of having a tooth made of the element with atomic number 79, but the concept *has a tooth made of gold* is not the concept *has a tooth made of the element with atomic number 79*. Similarly, for any mental state S, the concept *is in S and such that gold is the element with atomic number 79* is necessarily coextensive with the concept *is in S*, so they are both concepts of S.

Of the conceptual and metaphysical contrasts, neither immediately entails the other. If the concept *knows* is mental while the concept *believes truly* is not, then it follows immediately that knowing is a mental state, but it does not follow immediately that believing truly is not a mental state, for perhaps there could also be a mental concept of the state of believing truly. Thus the conceptual contrast does not immediately entail the metaphysical contrast. If knowing is a mental state and believing truly is not a mental state, then it follows immediately that the concept *believes truly* is not mental, but it does not follow immediately that the concept *knows* is mental, for perhaps there could be a different concept of the state of knowing which was mental. Thus the metaphysical contrast does not immediately entail the conceptual contrast. Nevertheless, it is hard to see why someone should accept one contrast without accepting the other. If the concept *believes truly* is non-mental, its imagined necessary coextensiveness with a mental concept would be a bizarre metaphysical coincidence. If the concept *knows* were a non-mental concept of a mental state, its necessary coextensiveness with a mental concept would be an equally bizarre metaphysical coincidence. In practice, sloppily ignoring the distinction between the metaphysical and conceptual contrasts is unlikely to do very much harm. Nevertheless, it is safer not to ignore the distinction.

The concept *believes truly* is not a mental concept of a state. If the concept C is the conjunction of the concepts $C_1, \ldots, C_n$, then C is mental if and only if each C_i is mental. For example, the conjunctive concept *is sad and such that gold is the element with atomic number 79* is non-mental, simply because it has the non-mental conjunct *is such that gold is the element with atomic number 79*, although it is a concept of the state of sadness. Even a logically redundant non-mental component concept would make C a non-mental concept, although it would then be logically equivalent to a mental concept. By contrast, non-mental concepts in the content clause of an attitude ascription do not make the concept expressed non-mental; the concept *believes that there are numbers* can be mental even if the concept *number* is not. At least, all that is so in a reasonable sense of 'mental', which one might express as 'purely mental'. Now the concept *believed truly* is the conjunction of the concepts *believed* and *true*. The conjunct *true* is not mental, for it makes no reference to a subject. Therefore, the concept *believed truly* is non-mental. Similarly, the concept *believes truly* of subjects rather than propositions is non-mental. The metaphysical and conceptual contrasts turn on whether knowing is a mental state, and on whether *knows* is a mental concept.

Just as the concept *believes truly* is non-mental, so for a similar reason is the concept *has a justified true belief*. Indeed, such an argument applies to any of the concepts with which the concept *knows* is equated by conjunctive analyses of the standard kind. The argument can be generalized to analyses formed using logical connectives other than conjunction. It would not apply if those simpler concepts were all mental, but analyses of the concept *knows* of the standard kind always involve irredundant non-mental constituents, in particular the concept *true*. Consequently, the analysing concept is non-mental: that is, not purely mental. Given that the concept *knows* is mental, every analysis of it of the standard kind is therefore incorrect as a claim of concept identity, for the analysing concept is distinct from the concept to be analysed.

If a non-mental concept were necessarily coextensive with the mental concept *knows*, they would be concepts of the same mental state. The present account does not strictly entail that no analysis of the traditional kind provides correct necessary and sufficient conditions for knowing. But once we accept that the concept *knows* is not a complex concept of the kind traditionally envisaged, what reason have we to expect any such complex concept even to provide necessary and sufficient conditions for knowing?

Experience confirms inductively what the present account implies, that no analysis of the concept *knows* of the standard kind is correct. Indeed, the candidate concepts turn out to be not merely distinct from, but not even necessarily coextensive with, the target concept. Since Gettier refuted the traditional analysis of *knows* as *has a justified true belief* in 1963, a succession of increasingly complex analyses have been overturned by increasingly complex counterexamples, which is just what the present view would have led one to expect.[2]

Even if some sufficiently complex analysis never succumbed to counterexamples, that would not entail the identity of the analysing concept

[2] See Shope 1983 for the history of a decade of research into the analysis of knowing after Gettier 1963; an equally complex book could be written on post-1983 developments. Not all this work aims to provide an analysis in the traditional sense; see Shope 1983: 34–44.

with the concept *knows*. Indeed, the equation of the concepts might well lead to more puzzlement rather than less. For knowing matters; the difference between knowing and not knowing is very important to us. Even unsophisticated curiosity is a desire to *know*. This importance would be hard to understand if the concept *knows* were the more or less ad hoc sprawl that analyses have had to become; why should we care so much about *that*?[3]

On quite general grounds, one would not expect the concept *knows* to have a non-trivial analysis in somehow more basic terms. Not all concepts have such analyses, on pain of infinite regress; the history of analytic philosophy suggests that those of most philosophical interest do not. 'Bachelor' is a peculiarity, not a prototype. Attempts to analyse the concepts *means* and *causes*, for example, have been no more successful than attempts to analyse the concept *knows*, succumbing to the same pattern of counterexamples and epicycles. The analysing concept does not merely fail to be the same as the concept to be analysed; it fails even to provide a necessary and sufficient condition for the latter. The pursuit of analyses is a degenerating research programme.[4]

We can easily describe simple languages in which no necessary and sufficient condition for knowing can be expressed without circularity. Many fragments of English have that property. Why should we expect English itself to be different? Once 'know' and cognate terms have been removed, what remains of our lexicon may be too impoverished to frame necessary and sufficient conditions for knowing.

The programme of analysis had its origin in great philosophical visions. Consider, for example, Russell's Principle of Acquaintance: '*Every proposition which we can understand must be composed wholly of constituents with which we are acquainted*' (Russell 1910–11, at

[3] Craig 1990a makes an interesting attempt to explain the point of the concept of knowledge in the light of the failure of analyses of the standard kind. However, on the present view it remains too close to the traditional programme, for it takes as its starting point our need for true beliefs about our environment (1990a: 11), as though this were somehow more basic than our need for knowledge of our environment. It is no reply that believing truly is as useful as knowing, for it is agreed that the starting point should be more specific than 'useful mental state'; why should it be specific in the manner of 'believing truly' rather than in that of 'knowing'? See also Chapter 3 for discussion of why we value knowledge more than mere true belief.

[4] For sophisticated but uncompelling defence of conceptual analysis see Jackson 1998 and Smith 1994: 29–56, 161–4. However, the kind of analysis they defend constitutes little threat to the claim that knowing is a mental state in every reasonable sense of the latter term. They provide no reason to suppose that the concept *knows* can be non-trivially analysed in any sense in which paradigmatic mental concepts cannot be, or that it is somehow posterior in the order of analysis to the concept *believes*. See also Fodor 1998 for a discussion of the demise of definition.

Salmon and Soames 1988: 23). Russell calls the principle 'the fundamental epistemological principle in the analysis of propositions containing descriptions'. There may well be a reading on which it is correct. However, when the principle is combined with Russell's extremely intimate conception of acquaintance, it forces analysis to go deeper than the surface constituents of the evidently intelligible propositions of science and common sense, for our acquaintance with those surface constituents is not perfectly intimate.[5] In such a context, the programme of analysis has a philosophical point. Now the philosophical visions which gave it a point are no longer serious options. Yet philosophers continued to pursue the programme long after the original motivation had gone. Correct deep analyses would doubtless still be interesting if they existed; what has gone is the reason to believe that they do exist.

While the general point is conceded, it might nevertheless be claimed that we have special reason to expect an analysis of *knows*. For we already have the necessary condition that what is known be true, and perhaps also believed; we might expect to reach a necessary and sufficient condition by adding whatever knowing has which believing truly may lack. But that expectation is based on a fallacy. If G is necessary for F, there need be no further condition H, specifiable independently of F, such that the conjunction of G and H is necessary and sufficient for F. Being coloured, for example, is necessary for being red, but if one seeks a further condition whose conjunction with being coloured is necessary and sufficient for being red, one finds only conditions specified in terms of 'red': being red; being red if coloured.

There are other examples of the same phenomenon. Although *x* is a parent of *y* only if *x* is an ancestor of *y*, it does not follow that we implicitly conceptualize parenthood as the conjunction of ancestry with whatever must be added to ancestry to yield parenthood, or even that ancestry is conceptually prior to parenthood. Rather, *x* is an ancestor of *y* if and only if a chain of parenthood runs from *x* to *y* (more formally: if and only if *x* belongs to every class containing all parents of *y* and all parents of its members). Thus parents of *y* are automatically ancestors of *y*. If anything, parenthood is conceptually prior to ancestry; we use the necessary and sufficient condition for ancestry in terms of parent-

[5] We must also assume Russell's conception of propositions as at the level of reference rather than sense. In effect, Evans 1982 combines the Principle of Acquaintance with a conception of acquaintance much less extreme than Russell's. Of course, Russell's extremism here is no mere extraneous dogma; it is an attempt to solve puzzles about the identity and non-existence of denotation in intentional contexts. Unfortunately, the cure is worse than the disease.

hood to explain why ancestry is necessary for parenthood.[6] Again, x is identical with y only if x weighs no more than y, but it does not follow that the concept *is identical with* is the conjunction of *weighs no more than* with whatever must be added to it to yield the former concept, or even that *weighs no more than* is prior to *is identical with*. In this case we explain the entailment by Leibniz's Law: if x is identical with y, whatever holds of x holds of y too, so since x weighs no more than x, x weighs no more than y. We grasp Leibniz's Law without considering all its instances. In principle one could grasp it before having acquired any concept of weight. Necessary conditions need not be conjuncts of necessary and sufficient conditions in any non-trivial sense.

More generally, the existence of conceptual connections is a bad reason to postulate an analysis of a concept to explain them. For example, the axiom of extensionality says that sets with the same members are identical; it has as good a claim to conceptual truth as the proposition that knowledge entails belief. Nevertheless, the axiom is not explained by an analysis of the concept *set*, if an analysis provides a non-circular statement of necessary and sufficient conditions.

The working hypothesis should be that the concept *knows* cannot be analysed into more basic concepts.[7] But to say that is not to say that no reflective understanding of it is possible.

1.4 KNOWING AS THE MOST GENERAL FACTIVE MENTAL STATE

Knowing does not factorize as standard analyses require. Nevertheless, a modest positive account of the concept can be given, one that is not an analysis of it in the traditional sense. The one sketched below will appear thin by comparison with standard analyses. That may not be a vice. Indeed, its thinness will clarify the importance of the concept as more complex accounts do not.

[6] As noted in the Introduction, we cannot define 'x is a parent of y' by 'x is an ancestor of y and x is not an ancestor of an ancestor of y'.

[7] A further ground for suspicion of analyses of the concept *knows* in terms of the concept *believes* is that they seem to imply that the latter concept is acquired before the former. Data on child development suggest, if anything, the reverse order (see Perner 1993: 145–203 for discussion of relevant work). Crudely: children understand ignorance before they understand error. Naturally, the data can be interpreted in various ways, and their bearing on the order of analysis depends on subtle issues in the theory of concepts.

The main idea is simple. A propositional attitude is factive if and only if, necessarily, one has it only to truths. Examples include the attitudes of seeing, knowing, and remembering. Not all factive attitudes constitute states; forgetting is a process. Call those attitudes which do constitute states *stative*. The proposal is that knowing is the most general factive stative attitude, that which one has to a proposition if one has any factive stative attitude to it at all. Apparent counterexamples to this conjecture are discussed below. The point of the conjecture is to illuminate the central role of the concept of knowing in our thought. It matters to us because factive stative attitudes matter to us.

To picture the proposal, compare the state of knowing with the property of being coloured, the colour property which something has if it has any colour property at all. If something is coloured, then it has a more specific colour property; it is red or green or . . .. Although that specific colour may happen to lack a name in our language, we could always introduce such a name, perhaps pointing to the thing as a paradigm. We may say that being coloured is being red or green or . . ., if the list is understood as open-ended, and the concept *is coloured* is not identified with the disjunctive concept. One can grasp the concept *is coloured* without grasping the concept *is green*, therefore without grasping the disjunctive concept. Similarly, if one knows that A, then there is a specific way in which one knows; one can see or remember or . . . that A. Although that specific way may happen to lack a name in our language, we could always introduce such a name, perhaps pointing to the case as a paradigm. We may say that knowing that A is seeing or remembering or . . . that A, if the list is understood as open-ended, and the concept *knows* is not identified with the disjunctive concept. One can grasp the concept *knows* without grasping the concept *sees*, therefore without grasping the disjunctive concept.

We can give substance to the category of factive stative attitudes by describing its realization in a natural language. The characteristic expression of a factive stative attitude in language is a *factive mental state operator* (FMSO). Syntactically, an FMSO Φ has the combinatorial properties of a verb. Semantically, Φ is an unanalysable expression; that is, Φ is not synonymous with any complex expression whose meaning is composed of the meanings of its parts. A fortiori, Φ is not itself such an expression. Φ also meets three further conditions. For simplicity, they are stated here as conditions on an FMSO in English, although the general category is realized in other languages too. First, Φ typically takes as subject a term for something animate and as object a term consisting of 'that' followed by a sentence. Second, Φ is factive, in the sense that the form of inference from 'S Φs that A' to 'A' is deductively valid (the scrupulous will read quotation marks as corner quotes where appropriate). Third, 'S Φs that A' attributes a propositional attitude to S. On the present view, 'know' and 'remember' are typical FMSOs. Even with the following glosses, these remarks do not constitute a rigorous definition of 'FMSO', but they should make its extension moderately clear.

First, 'S Φs that A' is required to have 'A' as a deductive consequence, not as a mere cancellable presupposition. There is a use of the verb 'guess' on which 'S guessed that A' in some sense presupposes 'A'. However, this presupposition is cancellable by context, as the logical and linguistic propriety of the following sentences shows:

(1) I guessed incorrectly that he was guilty.

(2) I guessed that he was guilty and you guessed that he was innocent.

In contrast, the substitution of 'knew' for 'guessed' in (1) or (2) yields a contradiction. Incidentally, therefore, the implication from 'S does not know that A' to 'A' is not like that from 'S knows that A' to 'A', for only the former is cancellable. The following sentences are logically and linguistically proper:

(3) I did not know that he was guilty, for he was innocent.

(4) I did not know that he was guilty and you did not know that he was innocent.

In contrast, the substitution of 'knew' for 'did not know' in (3) or (4) yields a contradiction. If Φ is an FMSO, the implication from 'S Φs that A' to 'A' is not cancellable (see Grice 1989: 44–6 and 279–80 for cancellability and the presuppositions of 'know' respectively).

Second, FMSOs are stative: they are used to denote states, not processes. This distinction is linguistically marked by the impropriety of progressive tenses. Consider:

(5) She is proving that there are infinitely many primes.

(6) The shoes are hurting her.

*(7) She is knowing that there are infinitely many primes.

*(8) She is believing that there are infinitely many primes.

*(9) The shoes are fitting her.

Sentences (7)–(9) are deviant because 'know', 'believe', and 'fit' (on the relevant reading), unlike 'prove' and 'hurt', are stative. Of course, a verb may have both stative and non-stative readings, as in (10):

?(10) She is remembering that there are infinitely many primes.

On the salient reading of 'remember', (10) is deviant, but it might correctly be used to say that she is in the process of recalling that there are infinitely many primes (see Vendler 1967: 104 for more on the linguistic marks of statives).

Third, an FMSO ascribes an attitude to a proposition to the subject. Thus 'S Φs that A' entails 'S grasps the proposition that A'. To know that there are infinitely many primes, one must grasp the proposition that there are infinitely many primes, so 'know' passes the test. A verb with a sense like 'is responsible for its being the case that' would fail it. Thus, given that 'see' and 'remember' are FMSOs, one can see that Olga is playing chess or remember that she was playing chess only if one has a concept of chess. This is not to deny that one's perceptions and memories may have a content which one lacks the concepts to express; the point is just that the English constructions 'see that A' and 'remember that A' do not ascribe such content. Other constructions with those verbs behave differently; one does not need a concept of chess to see or remember Olga playing chess.

Fourth, an FMSO is semantically unanalysable. An artificial verb stipulated to mean the same as 'believe truly' would not be an FMSO. A semantically analysable expression has a more complex semantic role than that of simply denoting an attitude; its proper treatment would require an account of the meanings from which its meaning is composed. Thus it is best at this stage to concentrate on semantically unanalysable expressions. Verbs such as 'know' and 'remember' will be assumed to be semantically unanalysable. However, an FMSO is not required to be syntactically unanalysable. In English and some other languages, for example, the addition of the auxiliary 'can' often forms an FMSO (Vendler 1967: 104–6). Consider the following pair:

(11) She felt that the bone was broken.

(12) She could feel that the bone was broken.

The 'could' in (12) is not the 'could' of ability; (12) does not mean anything like:

(13) She had the ability to feel that the bone was broken.

A rough paraphrase of the salient reading of (11) would be: 'She intuitively believed that the bone was broken.' A rough paraphrase of the salient reading of (12) would be: 'She knew by the sense of touch that the bone was broken'. Sentence (12), unlike (11), entails 'The bone was broken'. Thus 'could feel' differs from 'felt' in two ways: it is factive, and it is perceptual. Neither of these differences would occur if 'could feel' were semantically analysable into 'could' and 'feel', for that would assimilate 'could feel' to 'had the ability to feel', which is neither factive nor perceptual. 'Could feel' is semantically fused. It is an FMSO; 'feel' is not.

'Hear' is like 'feel' in this respect. Consider:

(14) She heard that the volcano was erupting.

(15) She could hear that the volcano was erupting.

A rough paraphrase of the salient reading of (14) would be: 'She heard a report that the volcano was erupting.' A rough paraphrase of the salient reading of (15) would be: 'She knew by the sense of hearing that the volcano was erupting.' Sentence (15), unlike (14), entails 'The volcano was erupting'. Thus 'could hear' differs from 'heard' in two ways: it is factive, and it is more directly perceptual. Neither of these differences would occur if 'could hear' were semantically a compound of 'could' and 'hear'. 'Could hear' is an FMSO; 'hear' is not.

'Could see' differs from 'see' in only one of the two ways. Consider:

(16) She saw that the stock market had crashed.

(17) She could see that the stock market had crashed.

Both (16) and (17) entail 'The stock market had crashed'; there is no difference in factiveness. However, they are naturally read in such a way that (16) would be true and (17) false if she simply saw a newspaper report of the crash; (17) might be true if she saw investors lining the window ledges. In such cases, one could insert 'the news' before 'that' in (16) but not in (17)—not even when she has inferred the crash from newspaper reports of other events. In this way, 'could see' is more directly perceptual than 'saw'. This does not prevent both from being FMSOs.

The notion of an FMSO should by now be clear enough to be workable; it can be projected onto new cases. Moreover, it has been explained without essential reference to the notion of knowing, although 'know' is an example of an FMSO. It will now be proposed that 'know' has a special place in the class of FMSOs.

The proposal is that if Φ is any FMSO, then 'S Φs that A' entails 'S knows that A'. If you see that it is raining, then you know that it is raining. If you remember that it was raining, then you know that it was raining. Such entailments are plausible but not uncontroversial (see Unger 1972 and 1975: 158–83 for useful discussion).

It is sometimes alleged that one can perceive or remember that A without knowing that A, because one fails to believe or to be justified in believing that A. Other evidence may give one reason to think that one is only hallucinating what one is in fact perceiving, or only imagining what one is in fact remembering. One abandons the belief, or retains it without justification; either way, it is alleged, one fails to know (Steup 1992 is a recent example of such a view). However, such cases put more pressure on the link between knowing and believing or having justification than they do on the link between perceiving or remembering and knowing. If you really do see *that* it is raining, which is not simply to see the rain, then you know that it is raining; seeing that A is a way of knowing that A. You may not know that you see that it is raining, and consequently may not know that you know that it is raining, but neither condition is necessary for knowing that it is raining (see Chapter 5). Similarly, if you really do remember *that* it was raining, which is not simply to remember the rain, then you know that it was raining; remembering that A is a way of knowing that A. You may not know that you remember that it was raining, and consequently may not know that you know that it was raining, but neither condition is necessary for knowing that it is raining. But it is far from obvious that you do see or remember that it is or was raining in the cases at issue, and an account will now be suggested on which you do not.

There is a distinction between seeing that A and seeing a situation in which A. One difference is that only the former requires the perceiver to grasp the proposition that A. A normal observer in normal conditions who has no concept of chess can see a situation in which Olga is playing chess, by looking in the right direction, but cannot see *that* Olga is playing chess, because he does not know what he sees to be a situation in which Olga is playing chess. The present cases suggest another difference between the two notions of seeing. By looking in the right direction, you can see a situation in which it is raining. In the imagined case, moreover, you have enough concepts to grasp the proposition that it is raining. Nevertheless, you cannot see *that* it is raining, precisely because you do not know what you see to be a situation in which it is raining (given the unfavourable evidence). On this account, the case is a counterexample to neither the claim that seeing implies knowing nor the claim that knowing implies believing.

Similarly, there is a distinction between remembering that A and remembering a situation in which A. One difference is that only the former requires the rememberer to grasp the proposition that A. Someone whose memory is functioning normally but who has no concept of chess can remember a situation in which Olga was playing chess, but cannot remember *that* Olga was playing chess, because he does not know what he remembers to be a situation in which Olga was playing chess. The present cases suggest another difference between the two notions of remembering. You can remember a situation in which it was raining. In the imagined case, moreover, you have enough concepts to grasp the proposition that it was raining. Nevertheless, you cannot remember *that* it was raining, precisely because you do not know what you remember to be a situation in which it was raining (given the unfavourable evidence). On this account, the case is a counterexample to neither the claim that remembering implies knowing nor the claim that knowing implies believing.

The discussion of FMSOs may be summarized in three principles:

(18) If Φ is an FMSO, from 'S Φs that A' one may infer 'A'.

(19) 'Know' is an FMSO.

(20) If Φ is an FMSO, from 'S Φs that A' one may infer 'S knows that A'.

The latter two principles characterize the concept of knowing uniquely, up to logical equivalence, in terms of the concept of an FMSO. For let 'schnow' be any term governed by (19′) and (20′), the results of substituting 'schnow' for 'know' in (19) and (20) respectively. By (19) and (20′), from 'S knows that A' one may infer 'S schnows that A'. Similarly, by (19′) and (20), from 'S schnows that A' one may infer 'S knows that A'. Thus 'schnow' is logically equivalent to 'know'. Note that this argument would fail if (20) held only for *most* FMSOs. In simple terms, 'know' is the most general FMSO, the one that applies if any FMSO at all applies.

In the material mode, the claim is that knowing is the most general stative propositional attitude such that, for all propositions *p*, necessarily if one has it to *p* then *p* is true. This is not quite to claim that, for all propositions *p*, knowing *p* is the most general mental state such that necessarily if one is in it then *p* is true. The latter claim fails for necessarily true propositions: every mental state is such that necessarily if one is in it then 5 + 7 = 12, but it does not follow that every mental state is sufficient for *knowing* that 5 + 7 = 12.

It is vital to this account of 'know' that 'believe truly' does not count as an FMSO. If it did, (20) would permit the invalid inference from 'S believes truly that A' to 'S knows that A'. The mental state is believing that A, not believing truly that A. To entail knowing, the mental state itself must be sufficient for truth. The condition of semantic unanalysability ensures that 'believe truly' does not count as an FMSO.

On this account, the importance of knowing to us becomes as intelligible as the importance of truth. Factive mental states are important to us as states whose essence includes a matching between mind and world, and knowing is important to us as the most general factive stative attitude. Of course, something needs to be said about the nature and significance of this matching, but that is a further problem. Someone who denied that the concept characterized by (18)–(20) is our concept *knows* might even think that it was more useful than the latter.

The states in question are general: different people can be in them at different times. No claim is made about the essences of their tokens; indeed, the idea of a token state is of doubtful coherence (Steward 1997: 105–34). With respect to general states, the claims of necessity are *de re*, not just *de dicto*. Given that 'knowing *p*' rigidly designates a mental state, the *de dicto* claim that the truth of *p* is necessary for knowing *p* implies the *de re* claim that for some mental state S the truth of *p* is necessary for S.

The account is explicitly not a decomposition of the concept *knows*; if 'know' were semantically analysable, it would not be an FMSO. It would certainly be quite implausible to claim that everyone who thinks that John knows that it is raining thereby thinks that John has the most general stative propositional attitude such that, for all propositions *p*, necessarily if one has it to *p* then *p* is true, to the proposition that it is raining. What, then, is the status of the account?

Consider an analogy. Identity is uniquely characterized, up to logical equivalence, by the principles of reflexivity and Leibniz's Law, just as knowing is uniquely characterized, up to logical equivalence, by (19) and (20). However, it would be quite implausible to claim that everyone who thinks that Istanbul is Constantinople thereby thinks that Istanbul bears to Constantinople the reflexive relation that obeys Leibniz's Law. The metalogical concepts used in formulating Leibniz's Law are far more sophisticated than the concepts we use in thinking that Istanbul is Constantinople. In order to have the concept *is* (of identity), one must somehow be disposed to reason according to Leibniz's Law, but that does not require one to have the metalogical concepts used in formulating Leibniz's Law. If it did, there would be an obvious danger of an infinite regress. Similarly, in order to have the concept *knows*, one must somehow be disposed to reason according to (18)–(20), but that does not require one to have the metalinguistic concepts used in formulating (18)–(20).

It is no straightforward matter to say what it is for a subject to be disposed to reason according to rules which the subject cannot formulate. Such a subject may even consciously reject the rules; philosophers who mistakenly deny Leibniz's Law do not thereby cease to understand the 'is' of identity. Nevertheless, some such notion does seem to be needed, independently of the account of knowing; the latter account can avail itself of that notion, whatever exactly it proves to be. The present account of knowing is consistent with the main features of a theory of concepts such as that of Peacocke 1992, on which an account of a concept gives necessary and sufficient conditions for possession of the concept without any need to decompose the concept itself. However, the account is not committed to any general programme of Peacocke's kind in the theory of concepts.

The present account of knowing makes no use of such concepts as *justified*, *caused*, and *reliable*. Yet knowing seems to be highly sensitive to such factors over wide ranges of cases. Any adequate account of knowing should enable one to understand these connections. This challenge is not limited to the present account: standard accounts of knowing in terms of justification must enable one to understand its sensitivity to causal factors, and standard accounts of knowing in terms of causal factors must enable one to understand its sensitivity to justification; none of these tasks is trivial.

One way for the present account to meet the challenge is by exploiting the metaphysics of states. For example, a form of the essentiality of origins may apply to states; a necessary condition of being in some states may be having entered them in specific ways. States of perceiving and remembering have this feature, requiring entry along a specific kind of causal path. Thus the importance of causal factors in many cases of knowing is quite consistent with this account. More obviously, having an inferential justification of a specific kind may be essential to being in some mental states; having a proof is clearly a factive mental state. Thus the importance of justification in many cases of knowing is equally consistent with this account. Of course, these remarks merely adumbrate a strategy, without carrying it out. Chapters 2 and 3 explore the connections between epistemology and the nature of mental states further. We can see epistemology as a branch of the philosophy of mind. If we try to leave epistemology out of the philosophy of mind, we arrive at a radically impoverished conception of the nature of mind.

1.5 KNOWING AND BELIEVING

The account of knowing above makes no essential mention of believing. Formally, it is consistent with many different accounts of the relation between the two concepts. Historically, however, the view of knowing as a mental state has been associated with the view that knowing entails *not* believing. Prichard is a case in point (1950: 86–8). On standard analyses of knowing, in contrast, knowing entails believing. On some intermediate views, knowing is consistent both with believing and with not believing. It is therefore natural to ask how far the present account of knowing constrains the relation between knowing and believing.

We have two schemas to consider:

(21) If S knows that A then S believes that A.

(22) If S knows that A then S does not believe that A.

If (21) is invalid, then the programme of analysing the concept *knows* as a conjunction of *believes* with *true* and other concepts is stillborn. Once the programme has been abandoned, (21) can be examined without prior need for its vindication.

The schema (22) is quite implausible. Whether I know that A on being told that A depends constitutively on whether my informant knew that A (amongst other factors). Whether I believe that A on being told that A does not depend constitutively on whether my informant knew that A; it would have to if knowing excluded believing. Of course, when one can describe someone as knowing that A, it is conversationally misleading simply to describe her as believing that A, but that is not to say that it is false. Not all believing is mere believing. We should reject (22).

The schema (21) does not sound *trivially* valid, as the schema 'If S knows that A then A' does. When the unconfident examinee, taking herself to be guessing, reliably gives correct dates as a result of forgotten history lessons, it is not an obvious misuse of English to classify her as knowing that the battle of Agincourt was in 1415 without believing that it was. But intuitions differ over such cases; it is not very clear whether she knows and not very clear whether she believes. In a case in which she was taught incorrect dates and repeats them with equal unconfidence, she is in an at least somewhat belief-like state, which she is also in when she was taught the correct dates. We have no clear counterexamples to (21) (see Radford 1966, Armstrong 1973: 138–49, and Shope 1983: 178–87 for further discussion of such cases).

There is a wide grammatical divergence between the verbs 'know' and 'believe' not suggestive of closely connected terms. For example, in a context in which I have predicted that it will rain, 'You know what I predicted' has a reading on which it is true if and only if you know that I predicted that it will rain, whereas 'You believe what I predicted' has no reading on which it is true if and only if you believe that I predicted that it will rain. There are many further grammatical differences between 'know' and 'believe' (see Austin 1946, Vendler 1972: 89–119, and Shope 1983: 171–8, 191–2). One explanation of such facts, proposed by Vendler, is that 'know' and 'believe' take different objects: what one knows is a fact, what one believes a proposition, where a fact is not a true proposition. A contingently true proposition, unlike a contingent fact, could have been false and still have existed. If so, then knowing is not a *propositional* attitude, and much of the terminology of this book might need revision, although the substance of the account would remain. Vendler's explanation makes it hard to see why (21) should be valid. However, it is not strictly inconsistent with the validity of (21), since 'that A' may refer to a fact in the antecedent and to a proposition in the consequent.

If 'that A' refers to a fact in the context 'S knows that A', then we might expect 'that A' to suffer reference failure when 'A' is false. Consequently, we might expect 'S knows that A' and 'S does not know that A' not to express propositions. But if 'A' is false, 'S knows that A' expresses a false proposition and 'S does not know that A' a true one. Perhaps we could treat 'that A' as elliptical for 'the fact that A' and analyse it by a Russellian theory of definite descriptions. The reference of 'fact that A' in the definite description is presumably determined by the proposition p expressed by 'A'; it is therefore some function f of p. Thus to know that A is to know the $f(p)$, and hence to stand in a complex relation expressed by 'know', 'the', and 'f' to the proposition expressed by 'A'. But then with only a slight change of meaning we could use the word 'know' for that complex relation to a proposition. Thus, even on a view like Vendler's, knowing would still involve a propositional attitude. However, it is very doubtful that there are any such things as facts other than true propositions (see Williamson 1999 for an argument). Moreover, the propriety of remarks like 'I always believed that you were a good friend; now I know it' and 'Long before I knew those things about you I believed them' suggest that 'believe' and 'know' do take the same kind of object. Vendler's account is not accepted here.

The present account of knowing might be thought inconsistent with the validity of (21), on the grounds that it provides no basis for a conceptual connection between believing and knowing. That would be too quick. Section 1.3 already noted that not every conceptually necessary condition is a conjunct of a conjunctive analysis. It is a mistake to assume that (21) is valid only if that connection is explicable by an analysis of *knows* in terms of *believes*. Consider an analogy: it may be a priori that being crimson is sufficient for being red, but that implication need not be explained by an analysis of one colour concept in terms of the other. One can grasp either concept without grasping the other, by being shown examples of its application and non-application. Neither concept relies on the other in demarcating conceptual space. Nevertheless, the area demarcated by one concept might be so safely within the area demarcated by the other that one could know by a priori reflection that the former is sufficient for the latter. Similarly, the area demarcated by the concept *knows* might be so safely within the area demarcated by the concept *believes* that one could know (21) by a priori reflection. That is quite consistent with, although not entailed by, the account of knowing in section 1.4.

An alternative proposal is to reverse the direction of analysis, and validate (21) by an analysis of *believes* in terms of *knows*. The simplest suggestion is that the concept *believes* is analysable as a disjunction of *knows* with other concepts. The word 'opine' will be used here as a term of art for the rest of the disjunction. On this analysis, one believes *p* if and only if one either knows *p* or opines *p*. Given that opining *p* is incompatible with knowing *p*, it follows that one opines *p* if and only if one believes *p* without knowing *p*. A similar view has been proposed by John McDowell (1982), building on the disjunctive account of perceptual experience developed by J. M. Hinton (1967 and 1973) and Paul Snowdon (1980–1 and 1990; see also Child 1994: 143–64, Dancy 1995, and Martin 1997). In McDowell's terminology, believing is not the highest common factor of knowing and opining. There is no such common factor. Rather, knowing and opining are radically different, mutually exclusive states, although instances of the latter are easily mistaken for instances of the former. Given a distinction between facts and true propositions, one could contrast knowing and opining somewhat as Vendler contrasts knowing and believing: to know is to be acquainted with a fact; to opine is to be acquainted with no more than a proposition. But the disjunctive conception does not require such an ontology of facts.

Not all those who advocate a disjunctive conception would claim that it provides a conceptual analysis. That claim faces difficulties additional to the generally dim prospects for conceptual analysis evoked in section 1.3. If the concept *believes* is the disjunction of *knows* and *opines*, then it must be possible to grasp the concept *opines* without previously grasping the concept *believes*. For otherwise, since grasping a disjunction involves grasping its disjuncts, it would be impossible to grasp the concept *opines* for the first time. Now 'opine' was introduced as a term of art; how is it to be explained? The natural explanation is that to opine a proposition *p* is to have a mere belief *p*, which is presumably to believe *p* without knowing *p*, but that explanation uses the concept *believes*. It does not permit one to grasp *opines* without already grasping *believes*. The explanation that to opine *p* is to be of the opinion *p* does no better, for 'be of the opinion' as ordinarily understood is just a rough synonym of 'believe'. In particular, once it is conceded—as it is by the disjunctive conception—that 'know' implies 'believe', little reason remains to deny that 'know' implies 'be of the opinion', too.

Can we explain 'opine' in terms of 'know'? A first attempt is this: one opines the proposition *p* if and only if one is in a state which one cannot discriminate from knowing *p*, in other words, a state which is, for all one knows, knowing *p*. That cannot be quite right, for if one cannot grasp the proposition *p* then one cannot discriminate one's state from knowing *p*; but one does not believe *p*, and therefore does not opine it. To avoid that problem, we can revise the definition thus: one opines *p* if and only if one has an attitude to the proposition *p* which one cannot discriminate from knowing, in other words, an attitude to *p* which is, for all one knows, knowing. However, that definition does not help a *disjunctive* analysis of believing. For if one knows *p*, then trivially one has an attitude to *p* which one cannot discriminate from knowing; one cannot discriminate something from itself. Thus the first disjunct, 'One knows *p*', entails the second disjunct, 'One opines *p*'. The whole disjunction would therefore be equivalent to its second disjunct, and the disjunctive form of the definiens would be a mere artefact of conceptual redundancy. To tack the qualification 'but does not know *p*' onto the end of the definition of 'opine' would make no significant difference, for since 'One either knows *p* or has an attitude to *p* which one cannot discriminate from knowing but does not know *p*' is still equivalent to 'One has an attitude to *p* which one cannot discriminate from knowing *p*', the disjunctive form would remain a mere artefact.

Alternatively, 'opine' might be explained as the disjunction of several more specific disjuncts, such as 'be under the illusion', 'be irrationally certain' and so on. However, it is very doubtful that, without using the concept *believes*, one could extend such a list to include all the different ways in which someone can believe without knowing. Those ways seem to be indefinitely various. How could one even specify, without using the concept *believes*, all the states in which someone can believe *p* falsely? If the list of disjuncts is open-ended, one could not grasp how to go on without realizing that one must list the ways in which someone can believe without knowing. Thus the explanation of 'opine' illicitly relies on a prior grasp of the concept *believes*.

The phenomenon just noted also threatens more metaphysical disjunctive accounts which do not attempt conceptual analysis, instead making their claims only about the underlying facts in virtue of which the concepts apply. Such an account of believing might deny that believing is itself a unified state, insisting that it is necessary but not a priori that one believes *p* if and only if one is in either the state of knowing *p* or the state of opining *p*. Since conceptual analysis is no longer in question, the replacement of 'opining' by 'merely believing' is not objectionable on grounds of circularity. The trouble is rather that there is no *more* reason to regard merely believing *p* as a unified mental state than to regard believing *p* as such. What unifies Gettier cases with cases of unjustified false belief is simply that in both, the subject believes without knowing; a good taxonomy of believing would not classify them together on the basis of some positive feature that excludes knowing. Moreover, it is hard to see how such a taxonomy could describe every species of believing without using the concept *believes*. But if a good taxonomy of believing does use the concept *believes*, that undermines the denial that believing is a unified state. Similar objections apply to disjunctive accounts of perception, appearance, and experience. For example, there is no reason to postulate a unified mental state equivalent to its appearing to one that A while one does not perceive that A.

A strictly disjunctive account of belief is not correct at either the conceptual or the metaphysical level. However, the disjunctive account was brought into play as a simple means to reconcile the account of knowing in section 1.4 with the supposed validity of (21) (knowing entails believing). There are other means to that end. A non-disjunctive analysis of *believes* might also validate (21). For example, (21) is a corollary of an analysis of *believes* itself on the lines of the definition of *opines* above: one believes *p* if and only if one has an attitude to the proposition *p* which one cannot discriminate from knowing, in other words, an attitude to *p* which is, for all one knows, knowing. That definition suggestively makes knowing central to the account of believing. One attraction of such an account is that it opens the prospect of explaining the difficulty, remarked by Hume, of believing *p* at will in terms of the difficulty of knowing *p* at will. The analysis is also consistent with the account of knowing in section 1.4.

Although that analysis provides a reasonable approximation to our concept *believes*, it does not fully capture the concept. It incorrectly classifies as believing that food is present a primitive creature which lacks any concept of knowing and merely desires that food is present; for all the creature knows, its attitude to the proposition that food is present is knowing. Equally incorrectly, the account classifies as not believing that there is a god someone who consciously takes a leap of faith, knowing that she does not know that there is a god. Both examples, however, are compatible with the variant idea that to believe *p* is to treat *p* as if one knew *p*—that is, to treat *p* in ways similar to the ways in which subjects treat propositions which they know. In particular, a factive propositional attitude to a proposition is characteristically associated with reliance on it as a premise in practical reasoning, for good functional reasons; such reliance is crucial to belief. A creature which lacks a concept of knowing can still treat a proposition in ways in which it treats propositions which it knows. The primitive creature does not treat the proposition that food is present like that when merely desiring that food is present; it does not use the proposition as a premise in practical reasoning. By contrast, the person who genuinely believes that there is a god by a leap of faith does rely on that premise in such reasoning. The unconfident examinee who tentatively gives *p* as an answer is little disposed to rely on *p* as a premise, and for that reason does not clearly believe *p*, but for the same reason does not clearly know *p*. Although a full-blown exact conceptual analysis of *believes* in terms of *knows* is too much to expect, we can still postulate a looser connection along these lines.

If believing *p* is, roughly, treating *p* as if one knew *p*, then knowing is in that sense central to believing. Knowledge sets the standard of appropriateness for belief. That does not imply that all cases of knowing are paradigmatic cases of believing, for one might know *p* while in a sense treating *p* as if one did not know *p*—that is, while treating *p* in ways untypical of those in which subjects treat what they know. Nevertheless, as a crude generalization, the further one is from knowing *p*, the less appropriate it is to believe *p*. Knowing is in that sense the best kind of believing. Mere believing is a kind of botched knowing.[8] In short, belief aims at knowledge (not just truth). These rather cryptic remarks will be developed in Chapters 9 and 10, which argue that knowledge is the evidential standard for the justification of belief.

Although the letter of disjunctive accounts has been rejected, the spirit may have been retained. For on the account in section 1.4, believing is not the highest common factor of knowing and mere believing, simply because it is not a factor of knowing at all (whether or not it is a necessary condition). Since that point is consistent with the claim that believing is common to knowing and mere believing, the claim is harmless. It no more makes the difference between knowing and mere believing extrinsic to a state than the point that continuity is common to straight and curved lines makes the difference between straight and curved extrinsic to a line. To know is not merely to believe while various other conditions are met; it is to be in a new kind of state, a factive one. What matters is not acceptance of a disjunctive account of believing but

[8] See also Peacocke 1999: 34.

rejection of a conjunctive account of knowing.[9] Furthermore, the claim that belief is what aims at knowledge is consonant with the suggestion in disjunctive accounts that illusion is somehow parasitic on veridical perception. Properly developed, the insight behind disjunctive theories leads to a non-conjunctive account of knowledge and a non-disjunctive account of belief.

While belief aims at knowledge, various mental processes aim at more specific factive mental states. Perception aims at perceiving that something is so; memory aims at remembering that something is so. Since knowing is the most general factive state, all such processes aim at kinds of knowledge. If a creature could not engage in such processes without some capacity for success, we may conjecture that nothing could have a mind without having a capacity for knowledge.

[9] Martin 1997: 88–90 questions whether a parallel account of perception and appearance will serve the purposes of naive realism, on the grounds that it does not entail the naive realist's distinctive claims about the phenomenology of perception. But a parallel account in terms of a factive mental state of conscious perceptual awareness may capture such claims.

2
Broadness

2.1 INTERNALISM AND EXTERNALISM

The thesis that knowing is a mental state faces a further series of challenges. They come from a picture of the mind known in current jargon as *internalism*, in a sense of the term more prevalent in the philosophy of mind than in epistemology.

When I attribute a mental state to you in ordinary language, the implications of my statement can easily outrun your boundaries. I say that you see paper; as every sceptic knows, you could be in the same internal state as someone who sees paper without seeing paper yourself: my statement is true only if paper is there before your eyes, outside you. In some sense, my statement is not purely about you. For theoretical purposes, would it not be more perspicuous to resolve the mixture into its underlying elements, by separating a statement purely about you from another purely about the environment external to you? After all, causation is local—no action at a distance—so does not the causal explanation of your actions require the isolation of what is local to you from background conditions on the environment? This resolution might amount to an analysis, giving a necessary and sufficient condition for the truth of my original statement. Alternatively, it might replace that statement without being equivalent to it, by doing its causal-explanatory work better. Either way, internal and external factors are separated. Such a picture is internalist.

In the present sense, internalists hold that one's mental states are determined by one's internal physical states; the mind is in the head. As a characterization of internalism, that is not fully general, for it neglects radically dualist versions, but the form of the argument below would be little changed if 'physical state' were replaced by 'phenomenal state', understood as designating states constitutively independent of the environment. For simplicity, we can focus on the currently most popular version of internalism.

Internalism provides a deeper motive for denials that knowing is a mental state. For since knowing is factive, whether one knows *p* constitutively depends on the state of one's external environment whenever the proposition *p* is about that environment. Consequently, whether one knows *p* is not determined by one's internal physical state. For example, whether one knows that it is raining is not determined by one's internal physical state, for it also depends on the weather. If it is not raining then one does not know that it is raining, whatever one's internal physical state. Thus, for the internalist, knowing is not a mental state. Jerry Fodor drew just such a conclusion from his formality condition, according to which mental states and processes defined over representations apply to them in virtue of the syntax of the representations: 'Since, on that assumption [that you can't know what's not the case], knowledge is involved with truth, and since truth is a semantic notion, it's going to follow that there can't be a psychology of *knowledge* (even if it is consonant with the formality condition to hope for a psychology of *belief*)' (1981: 228). By contrast, an externalist conception frees us to affirm that knowing is a mental state.

The issue ramifies. On the internalist picture, knowing is a metaphysical hybrid, a mixture of mental states with mind-independent conditions on the external world. Even Tyler Burge, who has done as much as anyone to develop an externalist understanding of the mental, writes that factive verbs like 'know', 'regret', 'realize', 'remember', 'foresee' and 'perceive' 'suggest an easy and clearcut distinction between the contribution of the individual subject and the objective, "veridical" contribution of the environment to making the verbs applicable' (1979: 85). Burge wisely adds in parentheses, 'Actually the matter becomes more complicated on reflection'. The internalist naturally tries to break the supposed mixture down into its elements, to analyse knowing in terms of believing, truth, and further factors. Even so-called externalist analyses of knowledge, where the further factors are causal or counterfactual, concede the internalist assumption that believing is somehow more basic than knowing. Thus internalism also provides a deeper motive for attempts to analyse knowing in terms of believing, truth, and other factors. We may call such attempts *the reductionist programme for knowledge*.

Internalists who regard knowing itself as complex do not thereby commit themselves to the same view of the concept *knows*. A simple concept might be defined by ostension of complex exemplars. Thus internalism motivates the reductionist programme for knowledge more strongly at the metaphysical than the conceptual level. The emphasis in this chapter will therefore be on metaphysical rather than conceptual issues.

We may assume that all attempts so far to carry out the reductionist programme for knowledge have failed. That suggests that it is misconceived. Its failure also suggests that internalism itself is misconceived, insofar as it motivates the reductionist programme. A conception of knowing that is thoroughly externalist in the present sense will dispense with the programme. On such a conception, as developed in the previous chapter, knowing is not a metaphysical hybrid, because it cannot be broken down into such elements.

Section 2.2 briefly illustrates the nature of the case for externalism, without attempting to state it in detail. The aim is rather to draw a comparison between more familiar disputes between internalists and externalists, over the contents of propositional attitudes, and the present dispute, over the attitudes to those contents, and to suggest that the case for externalism about mental attitudes is as good as the case for externalism about mental contents. The main target of criticism in sections 2.3 and 2.4 is the idea that there are good grounds for combining externalism about the contents of the attitudes with internalism about the attitudes themselves. One overall argumentative strategy is to show that objections to the involvement of factive attitudes in genuine mental states are sound only if corresponding objections to the involvement of broad contents in genuine mental states are also sound. For example, cases in which a difference in the external environment constitutes a difference in knowing hardly show that knowing is not a mental state, unless cases in which a difference in the external environment constitutes a difference in broad content show that believing a broad content is not a mental state. Externalism about factive mental attitudes is as well placed as externalism about mental content. Chapter 3 will state a deeper case for externalism on both fronts.

2.2 BROAD AND NARROW CONDITIONS

We can define the issues in more rigorous terms to address them more effectively. What exactly is the distinction between the internal and the external? The boundaries of the agent which our attributions of mental states outrun are spatio-temporal boundaries. The spatial boundary is naturally identified with that of the agent's body, although for present purposes it could just as well be identified with that of the brain (or the head). But only what goes on within the agent's body at the time of action counts as internal, for past bodily goings on are not local in the sense in which causation is supposed to be local. The internal will be identified with the total internal physical state of the agent at the relevant time, the external with the total physical state of the external environment.

Everything said here will be consistent with the mildly physicalist assumption that the internal and the external are jointly exhaustive as well as mutually exclusive, in the sense that the total internal physical state of the subject and the total physical state of the external environment jointly determine the total state of the world: no difference in the latter without a difference in at least one of the former.

Some terminology will help. A *case* is a possible total state of a system, the system consisting of an agent at a time paired with an external environment, which may of course contain other subjects. A case is like a possible world, but with a distinguished subject and time: a 'centred world' in the terminology of David Lewis (1979). Different cases can distinguish different subjects and times. Whatever is nomically possible counts as 'possible' in the relevant sense; whether anything else does can be left open for present purposes.

A *condition* obtains or fails to obtain in each case. Conditions are specified by 'that' clauses. The pronoun 'one' and the present tense in such clauses refer to the distinguished agent and time respectively. Thus the condition that one is happy obtains in a case α if and only if in α the agent of α is happy at the time of α.

A condition C *entails* a condition D if and only if for every case α, if C obtains in α then D obtains in α. The conditions C and D are identical if and only if for every case α, C obtains in α if and only if D obtains in α. Truth-functions of conditions are defined in the obvious way; for example, the conjunction of C and D obtains in α if and only if both C and D obtain in α. The criterion of identity for conditions ensures that such truth-functions have unique values.

A case α is *internally like* a case β if and only if the total internal physical state of the agent in α is exactly the same as the total internal physical state of the agent in β. A condition C is *narrow* if and only if for all cases α and β, if α is internally like β then C obtains in α if and only if C obtains in β. In other terminology, narrow conditions supervene on or are determined by internal physical state: no difference in whether they obtain without a difference in that state. C is *broad* if and only if it is not narrow. A state S is narrow if and only if the condition that one is in S is narrow; otherwise S is broad. Internalism is the claim that all purely mental states are narrow; externalism is the denial of internalism.

When we attribute mental states to each other in ordinary language, the conditions of which we speak are often broad. That one sees Naples, that one remembers Naples, that one keeps referring to Naples—all are broad conditions, because none obtains in cases in which one lacks even indirect causal connection with Naples, whereas one's internal physical state has no such necessary dependence on a city. Similarly, that one loves Mary and that one hates Mary are broad conditions, for they depend on a relation to the particular individual named.

The semantics of ascriptions of content to propositional attitudes in natural languages is a notorious source of broad conditions. In retrospect we can trace the idea back to Hilary Putnam (1973), as interpreted in the light of Burge's work, such as his 1979, 1986a, and 1986b. For example, the sentence 'One believes that there are tigers' expresses a broad condition. To check that it does, consider a counterfactual world like ours except that the only tiger-like creatures, although similar in appearance to tigers, are quite different in evolutionary ancestry and internal constitution. The differences are in respects about which ordinary non-zoologists are ignorant. Call the tiger-like creatures *schmigers*. Clearly, schmigers do not belong to the same species as tigers; they are not tigers. In the counterfactual world I have a *doppelgänger*, twin-TW, who is in exactly the same internal physical state as I am. I believe truly that there are tigers. I express my belief by saying 'There are tigers'. Twin-TW expresses his belief by saying 'There are tigers', too. If he believes that there are tigers then he is wrong, for in his circumstances there are no tigers; there are only schmigers. But twin-TW is no more mistaken on this matter than I am; both of us are ignorant rather than mistaken about those specific features that differentiate tigers from schmigers. Since twin-TW believes truly, he does not believe that there are tigers. Rather, his belief is true if and only if there are schmigers. Thus I differ from twin-TW in believing that there are tigers, even though we are in exactly the same internal physical state. John McDowell (1977) and Gareth Evans (1982) identify a similar phenomenon in relation to singular thoughts. I believe that this screen flickers; someone could be in exactly the same total internal physical state without believing that this screen flickers, because what (if anything) he believes flickers is not this screen but another with the same appearance in front of him. Thus the condition that one believes that this screen flickers is not broad. Similar arguments apply to a vast range of contents, and of attitudes to those contents. We may say derivatively that a content is broad if, for every attitude, the condition that one takes that attitude to that content is broad. Most contents ascribed in natural language are broad.

The internalist is obliged to concede that content ascriptions in natural languages express broad rather than narrow conditions, but nevertheless insists that they consequently fail to reflect the structure of the underlying facts. On this view, such ascriptions characterize the subject by reference to a mixture of genuinely mental states and conditions on the external environment. The challenge to such an internalist is to make good this claim by isolating a level of description of contentful attitudes that is both narrow and genuinely mental, not merely neurophysiological. If there is such a mixture of the internal and the external, it should be possible to separate out its constituents. The broadness of content ascriptions in natural languages shows that the required level of description does not simply lie to hand, but must be constructed; its effect is therefore to put the burden of proof on the internalist.

Parallel considerations apply to internalism about the attitudes themselves. Factive propositional attitudes are a source of blatantly broad conditions, whether or not their contents are broad. Even when the sentence 'One believes that A' does not express a broad condition, the conditions expressed by 'One knows that A', 'One sees that A', and 'One remembers that A' are almost always broad. While conceding this, the internalist nevertheless insists that such constructions fail to reflect the structure of the underlying facts. Factive constructions are held to characterize the subject by reference to a mixture of genuinely mental states and conditions on the external environment. As before, the challenge to the internalist is to make good this claim by isolating a level of description that is both narrow and genuinely mental. The effect of the broad natural language semantics is again to put the burden of proof on the internalist. On the view developed in Chapter 1, the factive states are as genuinely mental as any states are. The internalist disagrees, and tries to find a narrow non-factive attitude that exhausts the mental reality underlying the broad factive attitude. The next section examines such attempts; they all prove to be inadequate.

2.3 MENTAL DIFFERENCES BETWEEN KNOWING AND BELIEVING

The argument from internalism to the denial that knowing is a mental state can now be stated in more detail. First, assume that knowing is a mental state:

(1) For every proposition p, there is a mental state S such that in every case α, one is in S if and only if one knows p.

Given (1), any difference in knowing involves a difference in mental states. That is, (1) entails that knowing p *supervenes* on one's (total) mental state in this sense:

(2) For all propositions *p* and cases α and β, if one is in exactly the same mental state in α as in β, then in α one knows *p* if and only if in β one knows *p*.

The argument from (1) to (2) is immediate if one defines 'one is in exactly the same (total) mental state in α as in β' as 'for all mental states S, in α one is in S if and only if in β one is in S'. Whether, conversely, (2) entails (1) depends on whether what supervenes on one's mental state is itself a mental state, a question which need not be settled here. Statement (2) could also stand on its own, without such an analysis of the equivalence relation of exact sameness of mental state; then, unlike (1), it would involve no commitment to an ontology of mental states or any consequent problems in individuating states.

Whether or not it is derived from (1), (2) is commensurable with the internalist premise that one's mental state supervenes on one's physical state, in other words, that the condition that one is in a given mental state is narrow:

(3) For all cases α and β, if α is internally like β, then one is in exactly the same mental state in α as in β.

Together, (2) and (3) entail that knowing *p* supervenes on one's internal state, for supervenience is transitive:

(4) For all propositions *p* and cases α and β, if α is internally like β, then in α one knows *p* if and only if in β one knows *p*.

According to (4), the condition that one knows *p* is narrow. But (4) is uncontroversially false. Of two people in exactly the same internal physical state, one may know that it is raining while the other, as the result of an elaborate hoax, believes falsely that it is. One can know *p*, all but for the state of the environment, without knowing *p*, in the sense that one can be in exactly the same internal physical (not: mental) state as someone who knows *p* without oneself knowing *p*. Since internalists accept (3), they deny (2), the other premise from which (4) was deduced. Since (1) entails (2), they also reject (1).

Having denied that knowing is a mental state, internalists naturally seek to factorize it into mental and non-mental components. Since believing may appear to be determined by one's internal physical states, and therefore to qualify as a mental state by internalist lights, it is an obvious candidate constituent. The idea that the mental (or psychological) component of knowing is simply believing seems to be expressed in a remark by Stephen Stich, endorsed by Jaegwon Kim: 'what knowledge adds to belief is psychologically irrelevant' (Stich 1978: 574, quoted in

Kim 1993: 188; Kim 1993: 175–93 is a clear statement of the kind of view opposed here). Because believing is such an obvious candidate, even those who concede externalism about mental content may be inclined to internalism about the attitude of knowing, regarding it as a mixture of mental and non-mental elements.

In present terms, the claim that knowing *p* adds nothing mental to believing *p* comes to this:

(5) For all propositions *p* and cases α, if in α one believes *p* then in some case β one is in exactly the same mental state as in α and one knows *p*.

For if (5) is false, one can believe *p* while in a total mental state T incompatible with knowing *p*; but then the information that one knows *p* adds something mental to the information that one believes *p*, for it implies that one is in a total mental state other than T. Thus if knowing *p* adds nothing mental to believing *p*, then (5) holds. Conversely, if (5) holds, then knowing *p* imposes no constraints on one's mental state beyond those already imposed by believing *p*, so knowing *p* adds nothing mental to believing *p*.

Now (5) implies that knowing *p* is not a mental state, given that not knowing *p* is compatible with believing *p*. For if knowing *p* is a mental state, then anyone in exactly the same mental state as someone who knows *p* also knows *p*. More precisely, (1) formalizes the claim that knowing *p* is a mental state; (1) entails (2); (2) and (5) entail that in every case if one believes *p* then one knows *p*. Since not knowing *p* is compatible with believing *p*, knowing is a mental state only if (5) is false.

However, (5) fails, for reasons independent of internalism. One kind of case involves false propositions about the subject's own mental state. For example, let *p* be the proposition that someone is alert. Suppose that in case α one falsely believes that someone is alert solely on the basis of one's false first-person present-tense belief that one is alert. If in case β one is in exactly the same mental state as in α, then in β one also believes that someone is alert solely on the basis of one's first-person present tense belief that one is alert; since one's level of alertness is itself a feature of one's mental state, then in β one is not alert, so one's belief that one is alert is false; since that false belief is the only basis for one's belief that someone is alert, one does not know that someone is alert. Thus (5) fails. Counter-examples also occur when someone believes a necessarily false proposition. It is possible falsely to believe that 79 + 89 = 158; it is impossible to know that 79 + 89 = 158. These examples do not depend on any externalist assumptions about the contents of the beliefs.

Since those counterexamples involve false beliefs, the internalist might suppose the remedy to lie in the revised claim that knowing p adds nothing mental to believing p truly:

(6) For all propositions p and cases α, if in α one believes p truly then in some case β one is in exactly the same mental state as in α and one knows p.

Statement (6) implies that knowing p is not a mental state, given that not knowing p is compatible with believing p truly.

Even (6), however, is subject to counterexamples which are independent of externalism. Someone may believe truly that garlic is healthy to eat for reasons so confused and irrational as to be incompatible with knowing that garlic is healthy to eat; since his confusion and irrationality is an aspect of his mental state, no one could be in exactly the same mental state and know that garlic is healthy to eat. Attempts to rule out such cases will be plagued by notorious difficulties in stating a correct justification condition on knowledge (see Shope 1983: 45–118).

Even if (6) had been defensible, it would not have solved the original problem, which was, in internalist terms, to isolate the mental component of knowing, to say what mental state knowing adds nothing mental to. By specifying that the belief be true, (6) fails to do that. This latter objection is not met by a justification condition added to (6).

Believing p is in any case too unspecific a state to constitute the mental component of knowing p. Knowing p excludes: believing p solely for sufficiently confused and irrational reasons. The supposed narrow mental component of knowing p must include not just believing p but doing so without those kinds of confusion and irrationality. We must therefore consider the suggestion that the mental component of knowing is *rationally* believing (Fricker 1999). 'Rationally' here need not imply the ability to articulate reasons, but only the avoidance of irrationality; languageless animals and young children may still count as knowing. In place of 'rationally believes' we could also write 'has a justified belief'. Now if rationally believing is the mental component of knowing, then the latter adds nothing mental to the former:

(7) For all propositions p and cases α, if in α one rationally believes p then in some case β one is in exactly the same mental state as in α and one knows p.

Now (7) may also escape the original counterexamples to (5), for the internalist might count as not believing rationally the subject who believes that $79 + 89 = 158$ or that someone is alert solely on the basis of a false belief that he is alert. Moreover, (7) implies that knowing p is not a mental state, given that not knowing *p* is compatible with rationally believing *p*.

Perhaps (7) looks congenial to externalism about the contents of one's attitudes but not to externalism about one's attitudes to those contents. It is not. Suppose that it looks and sounds to me as though I see and hear a barking dog; I believe that a dog is barking on the basis of the argument 'That dog is barking; therefore, a dog is barking'. Unfortunately, I am the victim of an illusion, my demonstrative fails to refer, my premise sentence thereby fails to express a proposition, and my lack of a corresponding singular belief is a feature of my mental state, according to the content externalist. If I rationally believe that a dog is barking, then by (7) someone could be in exactly the same mental state as I actually am and know that a dog is barking. But that person, too, would lack a singular belief to serve as the premise of the inference, and would therefore not know that a dog is barking. Contrapositively, according to (7), I do not rationally believe that a dog is barking, even though there need be nothing internal wrong with my thought processes. Consequently, if the contents of beliefs depend like that on the external environment, then so too does the attitude of rational belief to a given content. In brief, (7) combined with content externalism makes rational belief an externalist mental attitude. If taking the externalist attitude of rational belief to a given content can contribute to one's mental state, why cannot taking the externalist attitude of knowledge to that content also contribute to one's mental state? The combination of (7) and content externalism makes the denial that knowing is a mental state ill-motivated.

My belief that a dog is barking may easily be true in the example, so to replace 'rationally' by 'rationally and truly' would gain nothing.

Indeed, (7) faces a further problem, one independent of content externalism. We could make (7) trivially true by *defining* 'in case α one rationally believes *p*' as 'in some case β one is in exactly the same mental state as in α and one knows *p*'. If knowing *p* entails believing *p* and believing *p* is a mental state, then such a definition would ensure that believing *p* was a necessary condition for rationally believing *p*. But it would neither isolate the mental component of knowing in independent terms nor provide any reason to suppose the mental component to fall short of knowing itself. If (7) is to give positive support to the hybrid conception of knowing as a mixture of mental and non-mental components, as (5) was supposed to do, then we should be able to grasp the relevant concept of rationality independently of grasping the concept of knowledge. Can we? Consider a case α in which one believes that ticket #666 will not win the lottery solely on the basis that its probability of winning is only one in a million. In any case β in which one is in the same mental state as in α one believes that ticket #666 will not win the lottery only on the same probabilistic grounds; thus in β one does not know that ticket #666 will not win the lottery. If one had known that the ticket would not win, one would not have bought it. Consequently, by (7), in α one does not rationally believe that the ticket will not win the lottery. But in α one's belief is not irrational in any obvious sense independent of considerations of knowledge. It is based on relevant reasons; the problem is just that they are not of a kind that would permit the belief to constitute knowledge. Chapter 8 will argue that considerations of rational belief depend on considerations of knowledge.

Given the failure of (5)–(7), someone might try to capture the idea that the difference between knowing and believing is not mental in the claim that *not* knowing *p* adds nothing mental to believing *p*. By analogy with (5), the claim is formalized thus:

(8) For all propositions *p* and cases α, if in α one believes *p* then in some case β one is in exactly the same mental state as in α and one does not know *p*.

For if (8) is false, someone can believe *p* while in a total mental state T incompatible with not knowing *p*; but then not knowing *p* adds something mental to believing *p*, as the former but not the latter is sufficient, given that one believes *p*, for being in a total mental state other than T. Thus if not knowing *p* adds nothing mental to believing *p*, then (8) holds. Conversely, if (8) holds, then not knowing *p* imposes no constraints on one's mental state beyond those already imposed by believing *p*, so not knowing *p* adds nothing mental to believing *p*.

Now (8) implies that knowing *p* is not a mental state, given that knowing *p* is compatible with believing *p*. For if knowing *p* is a mental state, then anyone in exactly the same mental state as someone who knows *p* also knows *p*. More precisely, (1) formalizes the claim that knowing *p* is a mental state; (1) entails (2); (2) and (8) entail that in every case if one believes *p* then one does not know *p*. Since knowing *p* is compatible with and perhaps entails believing *p*, knowing is a mental state only if (8) is false.

However, (8) is implausible even from an internalist perspective. For example, the proposition *p* may concern the subject's own mental state, or be a necessary truth. Internalists will classify direct awareness that one is in pain as a mental state, holding it to depend on nothing external. Presumably, being directly aware that one is in pain is sufficient for both knowing and believing that one is in pain. Thus if one is directly aware that one is in pain, one believes that one is in pain, and could not be in exactly the same mental state without being directly aware, and therefore knowing, that one is in pain. Consequently, (8) fails. Similarly, internalists will classify grasping a proof that 79 + 89 = 168 as a mental state, holding it to depend on nothing external. Presumably, they will also hold it to be sufficient for both knowing and believing that 79 + 89 = 168. Thus if one grasps a proof that 79 + 89 = 168, then one believes that 79 + 89 = 168, and could not be in exactly the same mental state without grasping the proof that 79 + 89 = 168, and therefore knowing that 79 + 89 = 168. Again, (8) fails. From a strongly externalist perspective, direct awareness that one is in pain may depend on something external; for example, it may depend on the use of a word in the language community as a whole to mean *pain* rather than something more specific that excludes one's current sensation. Similarly, grasping a proof that 79 + 89 = 168 may depend on the mathematical practice of the community as a whole. But for the full-blooded externalist, these external dependencies do not suggest that being directly aware that one is in pain and grasping a proof that 79 + 89 = 168 are not mental states. They can stand as counterexamples to (8). A defender of (8) would have to take an intermediate position, on which knowing *p* always has non-trivial external necessary conditions not constitutive of the subject's mental state, no matter how trivial the proposition *p*. Although this combination of claims is not obviously incoherent, it also has no obvious motivation; (8) lacks the independent plausibility to provide a good reason not to classify knowing as a mental state.

Since the subject rationally and truly believes *p* in the examples that are problematic for (8), qualifying 'believes' by 'rationally' or 'truly' in (8) would not help.

The supposed mental component of knowing short of knowing itself is a postulate of philosophical theory, not something provided by our understanding of the relations between knowing and believing. We have no good reason to accept the theory which makes that postulate. The internalist has no head start in the attempt to fence off the mental implications of knowing. That is not yet to say that the attempt cannot succeed; just that we have no reason independent of any case for internalism in general to expect it to succeed.

2.4 THE CAUSAL EFFICACY OF KNOWLEDGE

One motive for internalism is the combination of the idea that genuine states are causally efficacious with the idea that mental states are causally efficacious only if narrow. No action at a distance: causation is viewed as local, involving only narrow mental states. Since the property of judging that there is a tiger ahead is broad (because its content is broad), such an internalist denies that the property is causally efficacious, locating causal efficacy in properties that supervene on the subject's internal physical state. After all, they determine the subject's immediate physical movements. Similarly, the internalist will deny that the broad state of knowing that there is a dangerous animal ahead is causally efficacious, locating causal efficacy in the supposedly narrow state of believing that there is a dangerous animal ahead, the state which the knower shares with the victim of a sceptical scenario. For example, according to Harold Noonan (1993: 291–2), knowledge 'is best regarded not as a psychological state, but as a complex consisting of a psychological state (belief) plus certain external factors - not because its status as knowledge is causally irrelevant in action explanation, but because it does not have to be cited, as such, in the psychological explanation of action at all'.

Much needs to be probed and questioned in these internalist ideas. We should not assume that the notion of causal efficacy is clear, or derived from fundamental science, or known to apply only to local connections. Nevertheless, suspicion is legitimate of a purported mental state, reference to which never plays an essential role in causal explanation. In the case of broad contents, a standard externalist move is to argue that attributions of them do play an essential role in causal explanations whose explananda are themselves characterized in broad terms. For example, a hunter shoots a tiger while his counterfactual *doppelgänger* shoots a schmiger. These descriptions of the actions are not capricious, for they are the very ones under which the actions were intended, and therefore the ones to be used when we are trying to see how the actions made sense from the subjects' point of view. Since our explanandum is that the hunter shot the tiger, our explanans will naturally involve broad descriptions of him not true of his *doppelgänger*, such as 'He believed that shooting a tiger would make him popular'.

Similar considerations apply to the role of factive attitudes in causal explanation. Consider a causal explanation as simple as 'He dug up the treasure because he knew that it was buried under the tree and he wanted to get rich'. Note that the explanandum ('He dug up the treasure') makes reference to objects in the environment (the treasure) as well as to the subject's immediate physical movements. The internalist cannot substitute 'believe' for 'know' in the explanation without loss, for the revised explanans, unlike the original, does not entail that the treasure was where he believed it to be; the connection between explanans and explanandum is therefore weakened. The explanans does less to raise the probability of the explanandum. As usual, the internalist may react by substituting 'believe truly' for 'believe'. The new explanation is 'He dug up the treasure because he believed truly that it was buried under the tree and he wanted to get rich'. That may be deemed as satisfactory as the original explanation, although it sounds much less natural. However, even the substitution of 'believe truly' for 'know' sometimes involves explanatory loss.

A burglar spends all night ransacking a house, risking discovery by staying so long. We ask what features of the situation when he entered the house led to that result. A reasonable answer is that he knew that there was a diamond in the house. To say just that he believed truly that there was a diamond in the house would be to give a worse explanation, one whose explanans and explanandum are less closely connected. For one possibility consistent with the new explanans is that the burglar entered the house with a true belief that there was a diamond in it derived from false premises. For example, his only reason for believing that there was a diamond in the house might have been that someone told him that there was a diamond under the bed, when in fact the only diamond was in a drawer. He would then very likely have given up his true belief that there was a diamond in the house on discovering the falsity of his belief that there was a diamond under the bed, and abandoned the search. In contrast, if he *knew* that there was a diamond in the house, his knowledge was not essentially based on a false premise. Given suitable background conditions, the probability of his ransacking the house all night, conditional on his having entered it believing truly but not knowing that there was a diamond in it, will be lower than the probability of his ransacking it all night, conditional on his having entered it knowing that there was a diamond in it. It follows that the probability of his ransacking the house all night, conditional on his having entered it believing truly that there was a diamond in it, is lower than the probability of his ransacking it all night, conditional on his having entered it knowing that there was a diamond in it. In this case, the substitution of 'believe truly' for 'know' weakens the explanation, by lowering the probability of the explanandum conditional on the explanans. The substitution of 'believe' without 'truly' for 'know' would do even worse. The argument does not assume that lowering the probability of the explanandum conditional on the explanans strictly entails loss of explanatory power: just that it results in such a loss when, as here, there are no compensating gains.

One might be puzzled for a moment by the thought that, in the circumstances, the burglar's true belief *constituted* his knowledge. Were the effects not the same whatever one calls it? However, this thought does not address the original problem, which concerned the causal efficacy of a *general* state. Different people can share the state of knowing that there was a diamond in the house; this state cannot be equated with, since it is not necessary for, believing truly that there was a diamond in the house. No doubt the particular circumstances that in some sense realize the state in a given case can be described in many different ways; what matters is how relevant those descriptions are to an understanding of the effect in question. It emerged above that the description 'knows *p*' is sometimes more relevant than the description 'believes truly *p*'.

In order to *prove* that reference to states of knowing is essential to the power of a causal explanation, one would need to show that it could not be eliminated in favour of any combination of believing, truth, and so on. There are infinitely many potential substitutes which might be proposed. All that can be done here is to sketch a general strategy for dealing with them; we must not expect to prove that the strategy cannot fail. Given a potential substitute for 'knows', suppose that it does not provide a necessary and sufficient condition for knowing. One then constructs possible cases in which the failure of necessity or sufficiency makes a causal difference, making the proposed substitute not even causally equivalent to knowing. The potential substitute avoids this problem only if it does provide a necessary and sufficient condition for knowing. Thus the search for a substitute for knowing in causally explanatory contexts is forced to recapitulate the history of attempts to analyse knowing in terms of believing, truth, and so on, a history which shows no sign of ending in success.

For example, the substitution of 'believe truly without reliance on false lemmas' for 'know' can bring causal-explanatory loss. Variants of the previous case can be constructed in which the burglar enters the house believing truly that there is a diamond in it without reliance on false lemmas, yet fails to know in virtue of misleading evidence which he does not then possess, but may discover in the course of his search, in which case he will abandon the search. The argument for explanatory loss runs as before. Although knowing is not invulnerable to destruction by later evidence, its nature is to be robust in that respect. Stubbornness in one's beliefs, an irrational insensitivity to counterevidence, is a different kind of robustness; it cannot replace knowing in all causal-explanatory contexts, for the simple reason that those who know *p* often lack a stubborn belief in *p*. The burglar's beliefs need not be stubborn. Similarly, he need not feel certain of them; subjective certainty cannot always replace knowing. The same applies to believing truly on the best possible evidence, for the example can be so constructed that the burglar's evidence, although good, is not the best possible. When one works through enough examples of this kind, it becomes increasingly plausible that knowing can figure ineliminably in causal explanations. It is causally efficacious in its own right if any mental state is (see Pettit 1986 and Child 1994: 204–16 for related discussion).[1]

This chapter has stated a preliminary case for externalism about both mental contents and factive mental attitudes to those contents. The next chapter deepens the case and places it in a wider context.

[1] Hyman 1999 argues plausibly for another connection between knowledge and action: one knows that A if and only if one's reason for doing something can be that A. But it does not follow that one can *explain* knowing that A as being able to do things for the reason that A (as Hyman wishes to do). Someone in a Gettier case who believes truly that A without knowing that A cannot do X_1 for the reason that A, and cannot do X_2 for the reason that A, ... But a single failure to know explains all these incapacities. If the incapacities constituted the failure to know, the correlation between the incapacities would be an unexplained coincidence.

3
Primeness

3.1 PRIME AND COMPOSITE CONDITIONS

The previous chapter argued, in a preliminary way, that internalism is false and that, on the externalist alternative, knowing is a genuine mental state. This chapter deepens the critique of internalism. It argues on structural grounds that the envisaged separation of internal and external factors is impossible. Many ordinary mental states are not equivalent to conjunctions of something purely internal with something purely external. This inequivalence is essential to the causal-explanatory work which their attribution can do. The internal does not play the distinctive role in the explanation of action that internalism predicts. On an externalist understanding of mental states, knowing is a central exemplar.

Here is a sketch of an internalist line of thought:

> The causing of my present action is here and now. Only narrow conditions supervene on the here and now; so narrow conditions must play a privileged role in the causal explanation of action. If a causal explanation of action cites a broad mental condition, an underlying narrow condition must do the real work. We can isolate that narrow condition by subtracting from the broad mental condition the environmental accretions that make it broad. We can recover the original broad mental condition from the narrow condition by adding back those accretions.[1]

The internalist conceives the original broad mental condition as the conjunction of the narrow condition and a condition as purely external as the former is purely internal—for instance, the condition that one believes truly that it is raining as the conjunction of the narrow condition that one believes that it is raining and the environmental condition that it is raining.

[1] Terminology is borrowed from Jackson's useful survey (1996), although without attribution of the argument to him.

Let us state the matter more formally. Recall that a case α is *internally like* a case β if and only if the total internal physical state of the agent in α is exactly the same as the total internal physical state of the agent in β. A condition C is *narrow* if and only if for all cases α and β, if α is internally like β, then C obtains in α if and only if C obtains in β. C is *broad* if and only if it is not narrow. We can define external likeness on the model of internal likeness: a case α is *externally like* a case β if and only if the total physical state of the external environment in α is exactly the same as the total physical state of the external environment in β. A condition C is *environmental* if and only if for all cases α and β, if α is externally like β, then C obtains in α if and only if C obtains in β. In other terminology, environmental conditions supervene on or are determined by the physical state of the external environment. A condition C is *composite* if and only if it is the conjunction of some narrow condition D with some environmental condition E. As a special case, a narrow mental condition is trivially composite, for it is the conjunction of itself with the environmental condition that holds in all cases whatsoever. C is *prime* if and only if it is not composite. The line of thought that began with the here-and-nowness of causation led to the conclusion that mental conditions are composite.

That internalist line of thought is inconclusive, not least because it uses ill-defined notions of adding and subtracting conditions. The next section will show its conclusion to be false; many of the mental conditions which we attribute to each other in ordinary language are prime. It begins with a preliminary exploration of primeness, compositeness, and some related notions.

3.2 ARGUMENTS FOR PRIMENESS

A broad mental condition entails various narrow conditions. Indeed, there is a strongest narrow condition which it entails, that is, a narrow condition which it entails and which entails every such narrow condition. To see this, let *virtual*-C be the condition which obtains in a case α if and only if C obtains in some case internally like α. Virtual-C is narrow because internal likeness is transitive and symmetric. C entails virtual-C because internal likeness is reflexive. Moreover, virtual-C entails every narrow condition which C entails; for if C entails a narrow condition D, and virtual-C obtains in a case α, then C obtains in some case β internally like α, so D obtains in β (since C entails D), so D obtains in α (since D is narrow); hence virtual-C entails D. Thus virtual-C is the strongest narrow condition which C entails. When C is a broad mentalistic condition ascribed in natural language, internalists regard virtual-C as the purely mental reality underlying C. In particular, when C is the condition that one knows *p*, they are tempted to identify virtual-C with the condition that one believes *p*, or the condition that one rationally believes *p*. Section 2.3 argued that those identifications are incorrect.

We can define the dual notion of the weakest narrow condition that entails C by substituting 'every' for 'some' in the definition of 'virtual-C'; it obtains in a case α if and only if C obtains in every case internally like α. However, the resulting condition will usually be impossible when C is broad. For what case α does the condition that one believes that tigers growl obtain in every case internally like α? Virtual-C, the strongest narrow condition which C entails, is the condition of interest to the internalist.

Given a condition C, there is also a condition—call it *outward-C*—which stands to the external as virtual-C stands to the internal. Outward-C obtains in a case α if and only if C obtains in some case externally like α. Just as virtual-C is the strongest narrow condition which C entails, so outward-C is the strongest environmental condition which C entails.

We can now identify narrow and environmental conditions of which a given condition, if composite, is the conjunction: they are virtual-C and outward-C respectively. If C is any conjunction of narrow and environmental conditions at all, then it is the conjunction of virtual-C and environmental-C. We can prove that as follows. Let C be the conjunction D ∧ E of a narrow condition D and an environmental condition E. Since C entails D, virtual-C entails D; similarly, outward-C entails E. Thus the conjunction of virtual-C and outward-C entails D ∧ E, that is, C. Conversely, C entails the conjunction of virtual-C and outward-C, whether or not C is composite. Consequently, C, if composite, is the conjunction of virtual-C and outward-C. To argue that C is prime is in effect to argue that C can fail to obtain when both virtual-C and outward-C obtain.

How can we show that a condition C is prime? Suppose that C obtains in two cases α and β. Consider a case γ internally like α but externally like β; we may assume that such a case is possible because otherwise that interdependence of the internal and the external would itself undermine the idea that they can be separated (see section 3.3 for more on this assumption). Now suppose that C is the conjunction of a narrow condition D with an environmental condition E. Then C obtains in γ. For since C entails D, D obtains in α; since D is narrow, D also obtains in γ, which is internally like α. Similarly, since C entails E, E obtains in β; since E is environmental, E also obtains in γ, which is externally like β. Since C obtains whenever both D and E obtain, and they both obtain in γ, C obtains in γ, as required. Thus we can show that C is prime simply by exhibiting three cases α, β, and γ, where γ is internally like α and externally like β, and C obtains in α and β but not in γ. We shall see below how to do that for most ordinary mental conditions.

Conversely, the condition C is composite if no such triple of cases exists, for then C obtains in γ whenever both virtual-C and outward-C obtain in γ, so the conjunction of virtual-C and outward-C entails C, and the converse entailment is automatic. Thus it is necessary as well as sufficient for C to be prime that it obtains in two cases but not in a case internally like one and externally like the other.

A picture may help (Fig. 1). The horizontal axis represents total internal physical states; the vertical axis represents physical states of the external environment. The point representing γ is in the same position on the horizontal axis as the point representing α (because γ is internally like α) and in the same position on the vertical axis as the point representing β (because γ is externally like β).

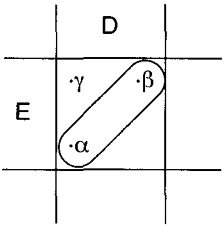

FIGURE 1

The area between the two vertical lines represents a narrow condition; the area between the two horizontal lines represents an environmental condition. The rectangle formed by their intersection represents a composite condition. The area enclosed by the curve represents a prime condition.

A structural analogy may also clarify what is going on. Suppose that a property P is the conjunction of a colour property Co and a shape property Sh, and that both a black sphere and a white cube have P. Then a black cube also has P: it has Co because it is the same colour as the black sphere, which has Co, and it has Sh because it is the same shape as the white cube, which has Sh. By contraposition, if a black sphere and a white cube have a property which a black cube lacks, then that property is not the conjunction of a colour property and a shape property.

Given a mental state S, how can we find three cases α, β, and γ of the required kind to show that the condition that one is in S is prime? We can construct them to a common pattern. We imagine circumstances in which S can be realized in just two ways, which need not be mutually exclusive. One is in S if and only if one is in S in either way 1 or way 2. Each way involves a channel with an internal and an external part; one is in S in way *i* if and only if both the internal and the external parts of way *i* are open (at this level of simplification, we may treat the condition that one is in S in way *i* as composite). In case α, both the internal and the external parts of way 1 are open but neither the internal nor the external part of way 2 is open; thus one is in S in way 1 although not in way 2; therefore one is in S. Case β reverses the two ways in status. In β, neither the internal nor the external part of way 1 is open but both the internal and the external parts of way 2 are open; thus one is in S in way 2 although not in way 1; therefore one is in S. In case γ, the internal part of way 1 but not the internal part of way 2 is open, because γ is internally like α, and the external part of way 2 but not the external part of way 1 is open, because γ is externally like β; thus one is in S in neither way 1 (because its external part is not open) nor in way 2 (because its internal part is not open); therefore one is not in S. The relations between α, β, and γ ensure that the condition that one is in S is prime. This structure can be represented diagrammatically (Fig. 2).

Case	Way	Internal	External	Joint	Result
α	1	✔	✔	✔	✔
	2	✘	✘	✘	
β	1	✘	✘	✘	✔
	2	✔	✔	✔	
γ	1	✔	✘	✘	✘
	2	✘	✔	✘	

FIGURE 2

Thus, as a first example, we can argue that the condition that one sees water is prime. Let α be a case in which one sees water normally with one's right eye. One's left eye receives light rays that by chance are like those it would receive from water, but they are all emitted by a waterless device just in front of that eye; however, a head injury prevents further processing of input from one's left eye. Let β be a case which differs from α by reversing the roles of the two eyes. In β, one sees water normally with one's left eye. One's right eye receives light rays that by chance are like those it would receive from water, but they are all emitted by a waterless device just in front of that eye; however, a head injury prevents further processing of input from one's right eye. Now consider a case γ internally like α and externally like β. In γ, a head injury prevents further processing of input from one's left eye, because it does so in α, and γ is internally like α. Equally, in γ, one's right eye does not receive light rays from water, because it does not do so in β, and γ is externally like β. Thus, in γ, neither eye both receives light rays from water and has its input to the brain subject to further processing. Consequently, in γ, one does not see water. Yet, in α and β, one does see water. By the earlier argument, the condition that one sees water is prime. Obviously, for almost any *x*, the example can be modified to show that the condition that one sees *x* is prime; it is not the conjunction of a narrow condition and an environmental condition.

That the example exploits the binocularity of vision is inessential. We could make the same point by supposing that in α there is water on the right and gin (which looks just like water) on the left, and a brain lesion causes one visually to register only what is on the right. In β there is gin on the right and water on the left, and a brain lesion causes one visually to register only what is on the left; in the case γ internally like α and externally like β, there is gin on the right and water on the left (as in β), and the brain lesion causes one visually to register only what is on the right (as in α). Thus, given appropriate background conditions one sees water in α and β but not in γ. This example is consistent with monocular vision.

For an aural analogue, suppose that, in α, Mary emits sound waves only of frequency *f* while John emits sound waves only of frequency *g*, and a brain lesion causes one aurally to register sound waves only of frequency *f*. In β, John emits sound waves only of frequency *f* while Mary emits sound waves only of frequency *g*, and a brain lesion causes one aurally to register sound waves only of frequency *g*. In the case γ internally like α and externally like β, John emits sound waves only of frequency *f* while Mary emits sound waves only of frequency *g* (as in β), and a brain lesion causes one aurally to register sound waves only of frequency *f* (as in α). Thus, given appropriate background conditions one hears Mary in α and β but not in γ. Examples of this type can be constructed for the other senses.

We can demonstrate the primeness of the condition that one believes that this screen flickers by substituting this screen for water in the first example. For we can assume that in α and β one's belief that this screen flickers concerns this screen only in virtue of one's visual link to this screen; in γ, since one fails to see this screen, one lacks the belief. The point generalizes to other object-dependent contents and other propositional attitudes.

Consider the condition that one believes that tigers growl. Let α be a case in which tigers inhabit the mountains while schmigers (which appear just like tigers) inhabit the jungle; one remembers one's encounters with tigers in the mountains but totally forgets one's encounters with schmigers in the jungle. One believes that tigers growl; since one has no recollection of schmigers, one does not believe that schmigers growl. Let β be a case in which tigers inhabit the jungle while schmigers inhabit the mountains; one remembers one's encounters with tigers in the jungle but totally forgets one's encounters with schmigers in the mountains. One believes that tigers growl; since one has no recollection of schmigers, one does not believe that schmigers growl. Now consider a case γ internally like α and externally like β. In γ, tigers inhabit the jungle while schmigers inhabit the mountains; one remembers one's encounters with schmigers in the mountains but totally forgets one's encounters with tigers in the jungle. One believes that schmigers growl; since one has no recollection of tigers, one does not believe that tigers growl. Thus the condition that one believes that tigers growl is prime.

We can make the example more vivid by supposing that one believes propositions by storing sentences in a language of thought which express them in a belief box, although it should be clear that nothing in the underlying structure of the example really requires a language of thought or a belief box. In each case, encounters with animals of a suitable appearance in the mountains cause one in a standard way to use a word T1 in one's language of thought (if at all) as a natural kind term for those animals. Encounters with animals of a suitable appearance in the jungle cause one in a standard way to use a word T2 in one's language of thought (if at all) as a natural kind term for those animals. One's language of thought also has a word G that means *growl*. In α, T1 means *tigers* and has the appropriate causal connections with tigers; one believes that tigers growl because one stores the sentence T1G in one's belief box. T2 does not mean *tigers*, because it lacks the appropriate causal connections with tigers; one does not store the sentence T2G in one's belief box. β differs from α by reversing the roles of T1 and T2. In β, T2 means *tigers* and has the appropriate causal connections with tigers; one believes that tigers growl because one stores T2G in one's belief box. T1 does not mean *tigers*, because it lacks the appropriate causal connections with tigers; one does not store T1G in one's belief box. In γ, one does not store T2G in one's belief box, because one does not do so in α and γ is internally like α. Equally, in γ, T1G does not express the proposition that tigers growl, because T1 does not mean *tigers*, since T1 lacks the appropriate causal connections with tigers in β and γ is externally like β. Thus, in γ, neither T1G nor T2G both expresses the proposition that tigers growl and is stored in the belief box. We can legitimately assume that in none of the three cases does any sentence in the language of thought other than T1G and T2G express the proposition that tigers growl. Consequently, in γ, one stores no sentence which expresses the proposition that tigers growl in one's belief box; one therefore fails to believe that tigers growl. Yet, in α and β, one does believe that tigers growl. Again, the point generalizes to other externally individuated contents and other propositional attitudes.

We can argue that epistemic conditions are also prime. The previous example might even suffice, for in some cases α and β of the specified kind one *knows* that tigers growl; in α, one fails to know that tigers growl because one fails to believe that tigers growl. However, it is more illuminating to pick an example in which the belief is held constant and what varies is its epistemic status. Let α be a case in which one knows by testimony that the election was rigged; Smith tells one that the election was rigged, he is trustworthy, and one trusts him; Brown also tells one that the election was rigged, but he is not trustworthy, and one does not trust him. Let β be a case which differs from α by reversing the roles of Smith and Brown; in β, one knows by testimony that the election was rigged; Brown tells one that the election was rigged, he is trustworthy, and one trusts him; Smith also tells one that the election was rigged, but he is not trustworthy, and one does not trust him. Now consider a case γ internally like α and externally like β. In γ, one does not trust Brown, because one does not trust him in α, and γ is internally like α. Equally, in γ, Smith is not trustworthy, because he is not trustworthy in β, and γ is externally like β. Thus, in γ, neither Smith nor Brown is both trustworthy and trusted. We can legitimately assume that in none of the three cases does one have any other way of knowing that the election was rigged. Consequently, in γ, one does not know that the election was rigged. Yet, in α and β, one does know that the election was rigged. Thus the condition that one knows that the election was rigged is prime. Since the example does not turn on the specific content of the knowledge, it can be modified to show for almost any proposition p that the condition that one knows p is prime.

Endless examples can be constructed to the foregoing pattern. Henceforth, mental conditions will therefore be assumed to be characteristically prime. They are not conjunctions of narrow conditions and environmental conditions.

3.3 FREE RECOMBINATION

When we construct triples of cases α, β, and γ to establish the primeness of a mental condition, the possibility of a case like γ depends on the principle that given cases α and β, there is a case internally like α and externally like β. Call that principle *free recombination*. It allows us to treat the internal and the external in a sense as independent variables.

Free recombination is not wholly unproblematic. If the internal and the external are nomically connected, then γ might violate nomic constraints even though α and β do not. For example, if determinism holds and the external includes the past (as it must for the treatment of issues about reference), then the external nomically determines the internal. Although a nomically impossible case might not be metaphysically impossible, such a case would not show very much, for if mental conditions coincided with conjunctions of internal and external conditions across all nomically possible cases that would be a significant vindication of the internalist picture of the mind. Moreover, the internal and the external are constitutively interdependent in other physical ways too. They are supposed to cover mutually exclusive and jointly exhaustive spatial regions; thus the region occupied by one determines the region occupied by the other. When the spatiotemporal interface between the internal and the external is contoured differently in α and β, mismatches threaten the construction of γ. Variations in the physical state of one include variations in the shape of the region it occupies, and therefore constrain variations in the physical state of the other.[2]

Nevertheless, free recombination may still hold to a first approximation, just as colour and shape can be treated to a first approximation as independent properties of an object, even if its colour ultimately depends on its microscopic geometry. We might handle the interdependence of the internal and the external just noted by minor modifications of the framework, for example, by restricting what aspects of the past count as environmental. Furthermore, counterexamples to free recombination are really a problem for the attempt to separate the internal and external under attack in this chapter. It is therefore fair to assume recombination in criticizing that attempt.

Consider, for example, Jerry Fodor's claim, 'identity of causal powers has to be assessed *across* contexts, not *within* contexts' (1987: 35), which he uses in defence of an individualistic conception of the mental (he has subsequently changed his position in his 1994). The proposal is

[2] See also Susan Hurley's related critique of what she calls the Duplication Assumption in ch. 8 of her 1998.

roughly that individuals have the same causal powers if and only if for every context *c*, they can do the same things in *c*. Of course, an individual's causal powers depend on its internal state, which must therefore be held fixed while the context varies. Thus a more precise formulation of Fodor's proposal is that an individual in an internal state I in some context has the same causal powers as an individual in an internal state J in a possibly different context if and only if for every context *c*, an individual in I in *c* can do the same things as an individual in J in *c*. Suppose that in case α one is in internal state I_α and context c_α; in case β one is in internal state I_β in context c_β. Are one's causal powers the same in α and β? By Fodor's criterion, a necessary (but insufficient) condition for sameness is that one can do the same things in I_α in c_β as one can do in I_β in c_β. This comparison breaks down, independently of what one can do in any case, unless one can be in I_α in c_β. But to be in I_α in c_β is to be in a case γ internally like α and externally like β, for internal likeness is sameness in one's internal state and external likeness is sameness in one's context. Thus the relevant comparisons can be made only if free recombination holds. Where recombination fails, Fodor's test does not grant sameness of causal powers, no matter what one can do in any case. Suppose, for instance, that any case internally like α and externally like β is discounted because it would violate the nomic constraints relative to which causal powers are being assessed; then one would not count as having the same causal powers in α and β. The same result follows if a case internally like α and externally like β is impossible because the spatio-temporal interface between the internal and the external is contoured differently in α and β. Yet these grounds for withholding the verdict of sameness of causal powers do not even mention what one can do; intuitively, they are inadequate. Fodor's test serves its purpose only if free recombination holds. Since his test is implicit in the internalist picture of mental causation, that picture requires free recombination. Thus an argument against the internalist picture can legitimately assume free recombination, for without it the picture fails anyway.

The dialectical position is similar for more general doubts about the distinction between the internal and the external. We should not assume without argument that subatomic physics will embody locality principles of a kind that would guarantee a clearcut distinction between those features which contribute to internal physical states and those which do not. The arguments for primeness are not restricted to mental conditions; they apply to other sorts of physical condition too (see also section 3.7). We may therefore find unexpected difficulty in defining the initial set of unproblematically internal physical conditions. That would destabilize the very distinction between the internal and the external. Such destabilization is bad news for the internalist conception of the mental under attack in this chapter, for that conception depends on separating the contributions of narrow and environmental conditions, which makes only as much sense as the distinction between the internal and the external itself does. If the distinction is unclear, the externalist can still insist on the negative point that no clarification of it counts ordinary mental conditions as composite. The present conception of the mental does not require a clearcut distinction between the internal and the external; we merely grant such a conception to its internalist rivals for the sake of argument, and then show that, even so, mental conditions cannot be decomposed into narrow and environmental conditions as they envisage.

3.4 THE EXPLANATORY VALUE OF PRIME CONDITIONS

Do concepts of prime conditions serve any theoretical purpose? In the examples that demonstrate primeness, what is the point of classifying case γ separately from cases α and β? This section argues that we need concepts of prime conditions for the same reasons for which we need concepts of broad conditions generally.

Consider seeing. What is the point of classifying a case in which one sees water separately from a case in which one is in exactly the same internal physical state but sees only a mirage? The difference may not matter for one's action at the next instant, if the action is itself individuated by its internal physical nature. If it were individuated broadly, we should still be wondering about the point of broad individuation. But our interest is not confined to action at the next instant—if there is one; if time is dense, there is not. One is thirsty; how likely is one to be drinking soon? Likely enough, if one sees water. Much less likely, if what one sees is a mirage. Even if drinking is individuated narrowly, its explanation in terms of the earlier state of the system involves the presence of water in the environment, not just the earlier internal physical state of the agent. Concepts of broad mental conditions give us a better understanding of connections between present states and actions in the non-immediate future, because the connections involve interaction with the environment (see also Burge 1986b and Peacocke 1993 on broad explanations of action).

The need to think about connections between earlier mental states and later actions is largely a need to think about connections between mental states and actions separated by seconds, days, or years. The causal explanation of action is frequently concerned with the structure of the agent's deliberation. But deliberation frequently occurs some time before the moment of action. In deliberating, one assesses alternative courses of action in the light of one's beliefs and desires, decides which is best, and forms the intention to pursue it; one puts the intention into effect only when the time for action comes. How and whether one puts the intention into effect depend on one's interaction with the environment in the intervening period. At the moment of action, one may not even remember one's deliberations in any detail. To confine the explanation of action to the instant before action is to omit much of what makes action rational. Historical explanation is certainly not confined to the instant before action; 'Why did Napoleon invade Russia?' is not a question about his state an instant before the invasion began, after months of planning.[3] Moreover, most actions take time; one cannot instantaneously eat an apple, write a letter, or go for a walk. Extended actions involve complex interaction with the environment.

Could we analyse each action into basic physical actions, and then explain each basic action in terms of the agent's internal state at the preceding instant? Conjoining those proximal explanations of the basic actions would not yield a good explanation of the original non-basic action. Suppose, for example, that we wish to explain why someone went for a walk. Perhaps we can analyse the walk into a sequence of steps. But for each step, the proximal explanation of his taking it will not mention what explains why he went for a walk: that he desired exercise.

For the reasons for which we need concepts of broad conditions, we need concepts of prime conditions. The relevance of seeing water now to drinking soon is not exhausted by the agent's internal state and the presence of water. Before one can drink the water, one must get oneself to it. Typically, one will steer one's way by keeping the water in sight and making a complex series of adjustments to one's position in a feedback loop. The present coincidence of one's internal physical state with the state in which one would be if one saw the water from that perspective is not enough; the coincidence must continue until one reaches the water. The kind of causal relation in which one stands to something

[3] Consider also Putnam's example (1978: 42): 'Professor X is found stark naked in the girls' dormitory at 12 midnight. Explanation: (?) He was stark naked in the girls' dormitory at midnight $-\varepsilon$, and he could neither leave the dormitory nor put on his clothes by midnight without exceeding the speed of light. But (covering law:) nothing (no professor, anyhow) can travel faster than light.'

when one sees it often enables one to keep it in sight. By contrast, if the matching of internal state and external environment is mere coincidence, then there is no reason why it should continue. We can find just this contrast in the cases which demonstrated the primeness of the condition that one sees water. Other things being equal, one can keep the water in sight with one's right eye in case α and with one's left eye in case β; in case γ, one has no means of maintaining the match between internal state and external environment. The match might continue, but that would be good luck. Even if it does continue, that is not seeing; whether one sees now does not depend like that on what happens in the future. Thus the need to understand the connection between present states and action in the non-immediate future gives us reason to classify case γ separately from cases α and β. To classify in that way is to use a concept of a certain prime condition.

We should not expect such considerations to yield a *definition* of seeing. As already noted, attempts to provide non-circular necessary and sufficient conditions for ordinary concepts have a miserable record of failure. The concept of seeing is the resultant of very many forces. The forces considered here make the concept of seeing a concept of a prime condition; they need not determine that condition uniquely.

The argument generalizes to other prime conditions. Sometimes, the content of a propositional attitude depends on a perceptual link to an object, as when one believes that this screen flickers; the foregoing considerations apply immediately. At other times, our thought of an individual or kind depends on a causal link independent of present perception. Nevertheless, such a link is typically *renewable*: it enables us to have further causal interactions with the individual or kind, or at least further causal dependencies on it if it no longer exists, for example, by finding out more about it and acting on that information—none of which implies that reference could be defined in terms of renewable causal links. Of the cases that demonstrated the primeness of the condition that one believes that tigers growl, such renewability is likely to be available in α and β but not in γ. In each case, one can test one's attribution of growling by going to the places where (some of) the creatures to which one attributes it were encountered; that takes one to the habitat of tigers in α and β and to the habitat of schmigers in γ. In encountering them there, one renews the causal link. If tigers really do growl while schmigers do not, one's belief is likely to survive in α and β but not in γ (we may suppose that in each case the tigers or schmigers were not growling when encountered but looked disposed to growl). Thus the need to understand the connection between present states and action in the non-immediate future gives us reason to use concepts of conditions that are prime because the content constitutively depends on a causal link with an individual or kind.

We can extend the argument to knowledge in more detail, taking a hint from Plato. In the *Meno* (97A–98A), Socrates raises a question about the value of knowledge. Knowing that this road goes to Larissa is useful to you if you want to go to Larissa. But merely believing truly that this road goes to Larissa seems to be equally useful to you, for you will get there just the same. Why should we value knowledge more than mere true belief? Socrates responds by a comparison with the statues of Daedalus, which run away unless they are tethered. True beliefs are liable to be lost, unless they are so anchored that they constitute knowledge.[4]

What does Plato mean? Surely he recognized that mere true beliefs can be held with dogmatic confidence, and knowledge lost through forgetting. But belief can also be sensitive to evidence. One can lose a mere true belief by discovering the falsity of further beliefs on which it had been essentially based; quite often, the truth will out. One cannot lose knowledge that way, because a true belief essentially based on false beliefs does not constitute knowledge. For example, I might derive the true belief that this road goes to Larissa from the two false (but perhaps justified) beliefs that Larissa is due north and that this road goes due north; when dawn breaks in an unexpected quarter and I realize that this road goes south, without having been given any reason to doubt that Larissa is due north, I abandon the belief that this road goes to Larissa. Since that true belief was essentially based on false beliefs, it did not constitute knowledge. The case is an obvious variation on Gettier's counterexamples to the analysis of knowledge as justified true belief.[5]

In other cases, a true belief not essentially based on false beliefs still fails to constitute knowledge, because misleading evidence against that true belief is rife in one's environment, although one happens to be unaware of it oneself. For example, I might correctly classify a dog by sight as friendly; in ordinary circumstances I might thereby come to know that it is friendly. However, this one behaves in ways which, if observed, would justify the false suspicion that it is hostile. So far it happens to have behaved like that only when my back was turned, and I have not yet formed any such suspicion. But my true belief that it is

[4] The passage is briefly discussed by Williams 1978: 38, Shope 1983: 12–13, and Craig 1990a: 7.

[5] The passage in the *Meno* is sometimes cited as a source of the view that knowledge is justified true belief; on this interpretation, Plato would be mistaken about the nature of the tether. The aim is not to elucidate his intentions here. See Fine 1992: 218–19 for further discussion and references.

friendly does not constitute knowledge, and could be lost at any moment. To know that it is friendly, I must not be surrounded by such misleading counterevidence, and my true belief must not be too vulnerable to this kind of overturning. The case is a variation on Harman's examples of the undermining of knowledge by evidence one does not possess (Harman 1973: 143–4 and 1980: 164–5; see also Goldman 1976: 772–3).

Present knowledge is less vulnerable than mere present true belief to *rational* undermining by future evidence, which is not to say that it is completely invulnerable to such undermining. If your cognitive faculties are in good order, the probability of your believing *p* tomorrow is greater conditional on your knowing *p* today than on your merely believing *p* truly today (that is, believing *p* truly without knowing *p*).[6] Consequently, the probability of your believing *p* tomorrow is greater conditional on your knowing *p* today than on your believing *p* truly today.[7] Of course, profoundly dogmatic beliefs which are impervious to future evidence and do not constitute knowledge may be even more likely to persist than beliefs that are rationally sensitive to future evidence and do constitute knowledge, but then the subject's cognitive faculties are not in good order. Since the difference between your present knowledge and your present true beliefs matters for predicting your future beliefs, it matters for predicting your future actions, because they will depend on your future beliefs.

The evidence which may undermine mere present belief needs time to emerge. As argued in section 2.3, the difference between knowledge and belief can make a present psychological difference; for instance, knowledge excludes various kinds of irrationality that belief does not. If C is the condition that one knows *p*, virtual-C can fail in several ways to be the condition that one believes *p*. However, the present argument concerns only delayed impact, not action at the 'next' instant. We do not value knowledge more than true belief for instant gratification.

We should not expect to define knowledge in terms of persistent true belief, still less in terms of subsequent action. What the argument does suggest is that when a condition stated in non-circular terms (belief, truth, justification, causation, . . .) fails to be necessary and sufficient for

[6] Knowing *p* is assumed to entail believing *p*; if not, the claim should be about the conditional probability of knowing or believing *p* tomorrow. Similar adjustments could be made throughout.

[7] The argument uses the easily established fact about conditional probabilities that if E entails D and $P[C|D \wedge \sim E] < P[C|E]$ and $P[E|D] < 1$, then $P[C|D] < P[C|E]$. Let C be the condition that you believe *p* tomorrow, D that you believe *p* truly today, and E that you know *p* today.

knowledge, that divergence will yield a divergence in implications for future action; the task of stating non-circularly a condition equivalent to knowledge with respect to implications for future action is no easier than the task of stating non-circularly a condition necessary and sufficient for knowledge. On the view defended in chapter 1, both tasks are impossible. Consequently, the mental state of knowing makes a distinctive contribution to the causal explanation of action.

These considerations apply to the cases that demonstrated the primeness of epistemic conditions. In α and β, one knows *p* by testimony, and one's belief in *p* is correspondingly stable; one has also been told *p* by someone untrustworthy whom one distrusts, but that does not make one's belief less stable, for it does not rest on that testimony. In γ, one believes *p* because one trusts the untrustworthy informant who tells one *p*; one distrusts the trustworthy informant who tells one *p*. One's belief is less stable, for in the long run one may recognize the untrustworthiness of the informant on whom one relies. That one may also recognize the trustworthiness of one's other informant is only partial compensation. The description of γ involves a specific threat to one's belief. The descriptions of α and β involve no such specific threat; although one might come to distrust the trustworthy informant, no reason has emerged why one should. On mildly cheerful assumptions about normal background conditions, there is at least some tendency for the truth on such matters to out, so one's belief in *p* is more stable in α and β than in γ. The need to understand the connection between present states and action in the non-immediate future gives us reason to use concepts of prime epistemic conditions.

3.5 THE VALUE OF GENERALITY

Some may nevertheless claim that prime conditions are theoretically redundant. For let α be any case in which a prime condition C obtains, and D and E respectively the strongest narrow condition and the strongest environmental condition which obtain in α. Then it is plausible that the conjunction D ∧ E entails C. For D completely specifies the internal physical state of the subject, and E completely specifies the physical state of the rest of the world, so, unless some physical relations fail to qualify as either internal or external (for example, for holistic reasons), D ∧ E completely determines the physical state of the world in α; and if the total state of the world supervenes on its total physical state, then D ∧ E entails C. Although the assumptions just made are not uncontroversial, let us allow them for the sake of argument. Thus, in any particular case, all the consequences we want of a prime condition follow from a compound condition obtaining in that case. Why should our best theory bother with the prime condition?

Let F be the condition that one subsequently performs a certain action. Suppose that F obtains in α; we want to know why. If the prime condition C makes F highly probable, given background conditions, we are tempted to cite C in explaining why F obtains. But D ∧ E presumably makes F certain, so why not cite D ∧ E rather than C? Doesn't it give the real causal explanation?

The answer to the would-be rhetorical questions is this. Our best theory is intended to capture significant generalizations. The action would have been performed in many cases other than α, in which D ∧ E does not obtain; D ∧ E is sufficient but nothing like necessary for F. A theory which relies on conditions like D ∧ E may leave uncaptured a significant generalization relating F to C. What has not been shown is that significant generalizations about prime conditions can be replaced by significant generalizations about compound conditions.

Good explanations have an appropriate generality. If one cites a sufficient condition for the condition to be explained, or one near enough so for the purpose in hand, the purported explanation can nevertheless fail because the condition to be explained would still have obtained in the same way even if the cited condition had not obtained. For example, one can explain why someone died by saying that he was run over by a bus; the explanation becomes worse, not better, if one specifies that the bus was red, for its colour had nothing to do with his death. If all metals have a certain property, one will be unhappy with attempts to explain why gold has it which cite properties of gold not shared by other metals. Again, if a condition obtains necessarily, then to explain why it obtains by deriving it from conditions which obtain only contingently is to miss its modal generality (as conventionalist explanations characteristically do).[8]

Many features of the maximally specific condition D ∧ E will be quite irrelevant to the obtaining of F. They will concern physical events that

[8] Consider also another of Putnam's examples (1978: 42): 'A peg (1 inch square) goes through a 1 inch square hole and not through a 1 inch round hole. Explanation: (?) The peg consists of such-and-such elementary particles in such-and-such a lattice arrangement. By computing all the trajectories we can get applying forces to the peg (subject to the constraint that the forces must not be so great as to distort the peg or the holes) in the fashion of the famous Laplacian super-mind, we determine that some trajectory takes the peg through the square hole, and no trajectories take it through the round hole. (Covering laws: the laws of physics.)'

form no part of the causal chain between the agent's initial mental state and the final performance of the action. The agent would have performed the action anyway, even if those features had been different. Their inclusion is a defect in the explanation. A highly specific account may constitute a good explanation of *how* something happened without constituting a good explanation of *why* it happened.[9]

Reductionist strategies of explanation risk providing bad explanations by citing highly specific conditions and thereby missing the generality of the conditions to be explained. Successful reductions involve no loss of generality. Something common to all genuine instances of the given phenomenon is identified in lower-level terms. The present argument does not undermine the explanatory value of those reductions. But that value is not shared by explanations which use no such generalization about the phenomenon, and merely provide—or rather gesture towards—a maximally specific description in lower-level terms of the particular case at hand.[10]

Defences of narrow content often treat the total state of the external environment ('circumstances', 'context') as one component of the favoured explanations of action, the other comprising attitudes to narrow contents. On the face of it, this is to give up generality across different states of the environment. Yet such accounts do not allow the internal component of the explanation to be similarly unarticulated; the condition that one has a certain attitude to a certain 'narrow content' is supposed to be a general one, obtaining in a range of different cases. Once generality is acknowledged as an explanatory virtue, the question arises whether it can best be achieved by explanations that factorize in the envisaged way.

We need concepts of prime conditions to achieve appropriate generality in explaining action. Consider again the cases that demonstrated primeness: α and β were mutually symmetric; the best explanation of the agent's subsequent actions might well generalize across α and β, citing a condition that obtains in both. But if it also obtained in γ, the explanation would be weakened, for then the cited condition would not rule out a range of cases in which the agent's subsequent actions in α and β are much less likely (see section 3.4). Thus the cited condition should obtain in α and β but not in γ, and therefore be prime.

Of course, generality is only one of many explanatory virtues. Some

[9] For example, the explanations envisaged by Jackson and Pettit 1995: 277 suffer from this loss of generality.

[10] See also Yablo 1992 and 1997 on proportionality between causes and effects, and Steward 1997: 192–7 on causal relations between facts.

purported explanations achieve spurious generality by using disjunctive concepts. For example, if someone was crying because she was bereaved, it does not improve the explanation to say that she was crying because she was bereaved or chopping onions. But ordinary mental concepts of prime conditions (such as the concept of seeing) are not disjunctive (see also section 1.5). To argue that such concepts do not express genuine common properties on grounds of explanatory uselessness would be viciously circular, for such concepts give generality to our explanations. Unless they are already assumed not to express common properties, nothing has been done to undermine their apparent explanatory usefulness.

When we explain why a condition C obtains by citing a prior condition D, the generality of our explanation varies inversely with P[C|~D], the probability of C conditional on ~D; in that sense, we lack generality to the degree to which D fails to be necessary for C.[11] The converse explanatory virtue is sufficiency, which varies with P[C|D], the probability of C conditional on D. We can combine these into a single explanatory virtue, the degree to which C is *correlated* with D. The higher the correlation, the better the answer to the question 'Does D obtain?' as a guide to the answer to the question 'Will C obtain?'; correlation is also a predictive virtue.

Correlation is itself only one of many explanatory virtues, but it is the one of present interest. A more rigorous framework for discussing it will be expounded, and then applied to the use of prime conditions in the causal explanation of action.

3.6 EXPLANATION AND CORRELATION COEFFICIENTS

In probability theory, the standard measure of correlation is the *correlation coefficient*, which takes values between +1 (perfect positive correlation) and −1 (perfect negative correlation; Appendix 1 gives technical details). Formally, this coefficient measures the correlation between random variables, which themselves take numerical values. For present purposes, what matter are correlations between conditions. We can

[11] The phrase 'degree of necessity' could equally well be associated with other measures, such as P[D|C]. An advantage of a probabilistic analysis is that it forces one to attend to such distinctions. For present purposes, the natural probabilities to consider are those conditional on the prior conditions D and ~D rather than those conditional on the outcome conditions C and ~C.

adapt the standard concept in a natural way to speak of the correlation coefficient of two conditions by associating each condition with its indicator random variable, which takes the value 1 when the condition obtains and 0 otherwise, and defining the correlation coefficient of two conditions as the correlation coefficient of their associated indicator random variables. The coefficient $\alpha[C,D]$ of correlation between the conditions C and D can then be calculated in terms of their probabilities and that of their conjunction. The result is a slightly unperspicuous formula:

$$\frac{P[C|D]-P[C]P[D]}{\surd(P[C](1-P[C])P[D](1-P[D]))}$$

The probabilities here are objective properties (chances) of the conditions, for the degree to which two conditions are correlated is an objective matter. But they are not single-case probabilities, for conditions are general, like properties; they can obtain in many actual cases. The relevant probability space comprises both actual and merely possible cases; they will be circumscribed by a set of background conditions, which vary with the explanatory context.

One can easily check that C and D are positively correlated ($\rho[C,D] > 0$) if and only if the probability $P[C|D]$ of C conditional on D exceeds the unconditional probability $P[C]$ of C (Appendix 1, proposition 2): one condition raises the probability of the other. The conditions are perfectly positively correlated ($\rho[C,D] = 1$) if and only if $P[C|D] = 1$ and $P[C|{\sim}D] = 0$ (Appendix 1, proposition 5): C is certain to obtain if D obtains and certain not to obtain if D does not obtain, so whether C obtains can be predicted with certainty on the basis of whether D obtains.

In explaining action, our concern is with imperfect positive correlations ($0 < \rho[C,D] < 1$). For example, the condition that one will perform a certain action in the future may be imperfectly positively correlated with both the condition that one presently knows some proposition and the condition that one believes that proposition truly (we can treat the relevant desires as background conditions). The question was: with which of the latter two conditions is the former condition better correlated? More generally, which of two conditions D and E, both positively correlated with C, is more highly correlated with C? It turns out that $\rho[C,D] \leq \rho[C,E]$ if and only if

$$(P[C|D]-P[C])(P[C]-P[C|{\sim}D]) \leq (P[C|E]-P[C])(P[C-P[C|{\sim}E])$$

(Appendix 1, proposition 3). Thus both the degree to which D helps us to predict C ($P[C|D]-P[C]$, the degree to which D raises the probability of C) and the degree to which ~D helps us to predict ~C ($P[C]-P[C|{\sim}D]$, the degree to which ~D lowers the probability of C) are relevant. In particular, if E raises the probability of C more than D does and ~E lowers the probability of C more than ~D does, then C is better correlated with E than with D (Appendix 1, proposition 4).

A completely schematic example may clarify the picture. We can take the example of knowledge, and develop the account in section 2.4 of its explanatory value. Similar points can be made about other prime conditions. Let C be the condition that one will perform a certain action, D the condition that one believes truly a proposition *p*, and E the condition that one knows *p*. The background conditions may include the agent's desires and other beliefs. Suppose that just three equiprobable possibilities have non-zero probability: one believes *p* truly and knows *p* and will perform the action; one believes *p* truly without knowing *p* and will not perform the action; one neither believes *p* truly nor knows *p* and will not perform the action. Hence $P[C \wedge D \wedge E] = P[{\sim}C \wedge D \wedge {\sim}E] = P[{\sim}C \wedge {\sim}D \wedge {\sim}E] = 1/3$. Thus it is certain that one will perform the action if and only if one knows *p*; $P[C|E] = 1$ and $P[C|{\sim}E] = 0$. The two conditions are perfectly positively correlated; $\rho[C,E] = 1$. It is also certain that one will perform the action only if one believes *p* truly, so $P[C|{\sim}D] = 0$. However, if one believes *p* truly, one may or may not perform the action, depending on whether one knows *p*; $P[C|D] = 1/2$. These two conditions are imperfectly positively correlated; calculation shows that $\rho[C,D] = 1/2$. Realistic examples have none of this simplicity. But since the correlation coefficient is a continuous function of the relevant probabilities, small enough changes in the latter make small changes in the former. Thus we can introduce some of the messy complexity of real life into the example and still have performing the action better correlated with knowing than with believing truly ($\rho[C,D] < \rho[C,E]$). In particular, this comparative ranking is consistent with non-zero probabilities for the possibilities that one performs the action without knowing *p*, with or without believing *p* truly ($C \wedge D \wedge {\sim}E$ and $C \wedge {\sim}D \wedge {\sim}E$) and that one fails to perform it while knowing *p* (${\sim}C \wedge D \wedge E$). Although in most realistic examples we cannot expect to calculate exact probabilities or correlation coefficients, such comparative rankings can still be plausible. In very crude terms, if the probability of performing the action conditional on knowing *p* exceeds the probability of performing it conditional on believing *p* truly without knowing *p* by much more than the latter exceeds the probability of performing it conditional on failing to believe *p* truly, then performing the action is more highly correlated with knowing *p* than with believing *p* truly. A precise statement of the principle would take into account the prior probabilities of knowing *p* and believing *p* truly.

Action is often more highly correlated with belief or with true belief than with knowledge. But not always. You see someone coming to your door; he is about to knock loudly. You are tempted not to reply. How would he react? You ask yourself, 'Does he know that I am in?' not, 'Does he believe that I am in?' If before knocking he does know that you are in, then he is unlikely to abandon his belief if you fail to reply; he will probably take offence. If before knocking he believes (truly) without knowing that you are in, then he is much more likely to abandon his belief if you fail to reply; he will probably not take offence. If before knocking he fails even to believe that you are in, then he is even less likely to take offence. Whether he would take offence is better predicted by whether he knows than by whether he believes. His taking offence is more highly correlated with knowing that you are in than with believing (truly) that you are in.

The point is not that knowing exceeds believing in implied degree of confidence; it need not. I know many things without being prepared to bet my house on them. The example works even if the degree of confidence required for belief is stipulated to be the same as the degree of confidence required for knowledge. The visitor who merely believes (truly) when he knocks that you are in may be exceedingly confident that you are in, but abandon the belief when to his astonishment you do not reply, for he is even more confident that if you do not reply you are not in.[12] But someone who knows that you are in has grounds that will not be undermined just by your failure to reply. Clearly, all this is a matter of probability; if your visitor merely believes that you are in, he *may* retain the belief when you do not reply, and take offence; equally, if he knows that you are in, he *may* abandon the belief in a fit of self-doubt when you do not reply. Nevertheless, the probabilities are often as indicated above.

Consider a variation on an example used in section 2.4, this time involving the attribution of beliefs to non-human animals. How long would we expect a fox to be willing to search for a rabbit in the wood before giving up, assuming initially (a) that the fox knows that there is a rabbit in the wood, or (b) that the fox believes truly that there is a rabbit in the wood? In (b) but not (a), the fox's initial true belief may fail to

[12] In probabilistic terms, if P_{new} is the result of updating P_{old} by conditionalization on the new evidence $\sim r$, then $P_{old}[i]$ can be arbitrarily high and $P_{new}[i]$ arbitrarily low, if $P[i|\sim r]$ is sufficiently low, which will require $P_{old}[r]$ to be high (let *i* be that you are in and *r* that you reply). Note that the relevant interpretation of 'P' here, in contrast with the rest of the chapter, is as degree of belief (credence), not as objective probability.

constitute knowledge because the true belief is essentially based on a false one, for instance, a false belief that there is a rabbit in a certain hole in the wood. When the fox discovers the falsity of that belief, the reason for the search disappears. That will not happen in (a), because a true belief essentially based on a false one does not constitute knowledge. Thus, given plausible background conditions, more persistence is to be expected in (a) than in (b). In many such cases, lengthy persistence is better explained by initial knowledge than by initial true belief.

Sometimes the predictive difference between knowledge and true belief is mediated by the cultural significance of knowledge. A notorious criminal may try to eliminate all those who know that he killed the policeman, because they are potential witnesses against him in court. He will not bother to eliminate those who merely believe truly that he did it, because their confidence that he did it, however great, is no threat to him, given the rules of forensic evidence. If we want to predict whether someone will soon be fleeing for her life, 'Does she know that he shot the policeman?' is a better question than 'Does she believe that he shot the policeman?'. It is better even than the question 'Does she believe that she knows that he shot the policeman?' when flight is contingent on the criminal's behaviour and he believes that the testimony of anyone who mistakenly believes themselves to know that he did it will not stand up in court. The danger is knowing too much, not believing too much.

We can also use a schematic example to reinforce the conclusion of section 3.5, by showing in detail how a very specific condition strictly sufficient for the condition to be explained can nevertheless fail to be highly correlated with it. Let C be the condition that one will perform a certain action, D the very specific condition obtaining in the case at hand (for example, completely determining the agent's internal physical state and the physical state of the environment), and E the condition that one knows *p*. The background conditions include the agent's desires. Assume that D is sufficient for both C and E, which leaves just five possibilities. Suppose that they have these probabilities:

$$\begin{aligned}
&P[C \wedge D \wedge E] = 1/10\\
&P[C \wedge {\sim}D \wedge E] = 3/10\\
&P[C \wedge {\sim}D \wedge {\sim}E] = 1/10\\
&P[{\sim}C \wedge {\sim}D \wedge E] = 1/10\\
&P[{\sim}C \wedge {\sim}D \wedge {\sim}E] = 4/10
\end{aligned}$$

Thus it is certain that one will perform the action if the specific condition obtains ($P[C|D] = 1$) and not certain that one will perform it if one knows *p* ($P[C|E] = 4/5$). However, one is much more likely to perform it if the specific condition does *not* obtain than if one does *not* know *p* ($P[C|{\sim}D] = 4/9 > 1/5 = P[C|{\sim}E]$). The latter disparity in favour of E more than compensates for the former disparity in favour of D; calculation shows that $\rho[C,D] = 1/3 < 3/5 = \rho[C,E]$. Performing the action is better correlated with knowing *p* than with the strictly sufficient specific condition.

3.7 PRIMENESS AND THE CAUSAL ORDER

A high correlation does not guarantee a direct causal connection. When the condition that one knows *p* is highly correlated with the condition that one will perform a certain action, the reason might be that both the knowledge and the action are effects of a common cause, without the knowledge causing the action. What would it be for the knowledge not to cause the action? Presumably, the condition that one knows would not be causally relevant in the right sense to the condition that one will perform the action. But then we should not focus on the former in explaining why the latter obtains. Does this seriously threaten the role of prime conditions in the explanation of action?

High correlations are an indispensable though fallible guide to causal structure. Where a high correlation misleads us into falsely postulating a causal connection, more detailed information about further correlations should correct our mistake. The high correlations between prime mental conditions and conditions on subsequent action constitute defeasible evidence for the causal effectiveness of the prime conditions. Higher correlations constituting a genuinely rival explanation would be needed to defeat that evidence.

Given deterministic laws, we might define a present condition D perfectly correlated with the condition C that one will perform the action, by stipulating that D obtains in a case α if and only if the total present state of the system (agent and environment) in α and the deterministic laws entail that one will perform the action. C and D obtain in exactly the same nomically possible cases. Thus, if the laws have probability 1, C and D are perfectly correlated (if $P[C] > 0$; otherwise $\rho[C,D]$ is ill-defined). D is not defined disjunctively. However, the definition of D unifies the cases in which D obtains by what happens later (the performance of the action), not by the present state of the system. In many contexts, such a correlation will not give us the kind of understanding we seek. It certainly does not give us what we need for purposes of prediction and control, but that is not quite the same thing. We seek a correlation between a condition given by a concept that unifies the cases in which it obtains in terms of the present state of the system and a condition given by a concept that unifies the cases in which it obtains in terms of the future state of the system; we are willing to sacrifice some degree of correlation in order to achieve such unification.

Even if we can replace the conditions conceived by folk psychology by conditions more highly correlated than they are with the condition that one subsequently performs the action, those new conditions will themselves be prime (as D is above), for reasons already indicated. Moreover, such explanations may well constitute refinements rather than refutations of the folk psychological explanations.

Discussions of broadness have tended to concentrate on intentional content. Since intentional content is a mental phenomenon, the causal efficacy of broad conditions, and specifically of prime conditions, can appear to require special pleading on behalf of the mental. It does not. The considerations of this chapter are not confined to the mental. For example, one can demonstrate in the style of section 3.2 the primeness of the condition that a ship is anchored to the seabed; it is not the conjunction of a condition on the internal physical state of the ship and a condition on the physical state of its external environment. Clearly, the condition that a ship is anchored to the seabed can be causally effective with respect to the ship's subsequent motion or rest. Primeness is no bar to causal efficacy. It derives its significance from our interest in causal-explanatory connections between states of objects and their subsequent behaviour (in the widest sense) after an interval long enough to permit intervening interaction with their environment. That is the normal case, not the exception, in causal explanation.

3.8 NON-CONJUNCTIVE DECOMPOSITIONS

The arguments in section 3.2 for the primeness of various mental conditions were not supposed to show that those conditions cannot be analysed somehow as functions of narrow and environmental conditions. A composite condition is the conjunction of a narrow condition with an environmental condition. How far do the problems identified in this chapter for conjunctive analyses generalize to analyses of other forms?

Conjunction is not the only truth-function. Disjunction is a simple alternative. Call a condition *non-trivial* if and only if it obtains in some cases but not in all. Then we can easily show, given free recombination, that the (inclusive) disjunction of a non-trivial narrow condition with a non-trivial environmental condition is always prime.[13] Thus an argument for the primeness of a mental condition does not automatically show that it is not such a disjunction.

Of course, given free recombination, we should not expect a mental condition to be the disjunction of a non-trivial narrow condition with a non-trivial environmental condition. For if it were, and the environmental condition obtained in a case β, then for any case α, some case γ would be internally like α and externally like β. Since the environmental condition would obtain in γ, the mental condition would too (because a disjunction is entailed by its disjuncts); and thus the mental condition would be consistent with any non-trivial narrow condition whatsoever (sawdust in the head, . . .), which is implausible. We could also argue that a mental condition C is not the disjunction of a narrow condition D and an environmental condition E by arguing that the contradictory condition ~C is prime, for if C were D v E then ~C would be ~(D v E), which is ~D ∧ ~E; since the contradictory of a narrow condition is itself narrow, and the contradictory of an environmental condition is itself environmental, ~D ∧ ~E is composite. But the possibility that a mental condition is a more complex function of narrow and environmental conditions cannot be dismissed so easily.

Suppose, for example, that a mental condition C involves some kind of matching between one's internal state and the state of the external environment, although it does not fix those states separately. Then a simple hypothesis would be that C is a possibly infinite disjunction $(D_1 \wedge E_1) \vee (D_2 \wedge E_2) \vee \ldots$, where $D_1, D_2, \ldots$ are narrow conditions, $E_1, E_2, \ldots$ are environmental conditions, and D_i matches E_i in the appropriate sense. Although each disjunct $D_i \wedge E_i$ is composite, disjunctions of composite conditions are not themselves usually composite. The required matching between internal and external states may occur in cases α and β separately without occurring in a case γ internally like α and externally like β; $D_i \wedge E_i$ and $D_j \wedge E_j$ may each entail matching while $D_i \wedge E_j$ does not. Equally, the matching may occur in γ without occurring in α or β, so the disjunction of composite conditions is not even the con-

[13] Proof: Suppose that the narrow condition D obtains in case δ but not in case δ^*, while the environmental condition E obtains in case ε but not in case ε^*. By free recombination, there are cases α, β, and γ, where: α is internally like δ^* and externally like ε; β is internally like δ and externally like ε^*; γ is internally like δ^* and externally like ε^*. Thus γ is internally like α and externally like β. Since E is environmental and obtains in ε, which is externally like α, E obtains in α; thus D v E obtains in α. Since D is narrow and obtains in δ, which is internally like β, D obtains in β; thus D v E obtains in β. But since E does not obtain in ε^*, which is externally like γ, E does not obtain in γ; since D does not obtain in δ^*, which is internally like γ, D does not obtain in γ; thus D v E does not obtain in γ. Therefore, D v E is composite.

tradictory of a composite condition. Thus, more realistically prime conditions can be constructed as quite simple truth-functions of narrow conditions and environmental conditions.

A less unsophisticated proposal is that the mental condition requires some causal relation between one's internal state and the matching state of the external environment. That would only strengthen the argument for primeness. Of course, causal relations to the environment are often conceived as themselves on the external side, in which case they could be subsumed under the environmental conditions; but since they also implicate their internal relata, that conception of them endangers free recombination. The causal relation is better conceived as bridging the internal and the external.

Since a prime condition may be a truth-function or some subtler function of narrow and environmental conditions, the arguments for the primeness of various mental conditions do not show that our concepts of those conditions cannot be analysed into concepts of narrow and environmental conditions. The arguments for unanalysability are different; as in section 1.3, they advert to the long history of failed analyses, the lack of any good reason to expect analysability, and the availability of an alternative understanding of the mental. Nevertheless, the arguments for primeness are needed to fix the role of the mental in the causal explanation of action. For even if a mental condition C were a disjunction $(D_1 \wedge E_1) \vee (D_2 \wedge E_2) \vee \ldots$ of conjunctions of non-trivial narrow conditions D_i with non-trivial matching environmental conditions E_i, it would not follow that C could be replaced in causal explanations by corresponding narrow and environmental conditions; a composite condition can be so replaced. Given free recombination, the strongest narrow and environmental conditions entailed by the disjunction are $D_1 \vee D_2 \vee \ldots$ and $E_1 \vee E_2 \vee \ldots$ respectively.[14] But if $(D_1 \wedge E_1) \vee (D_2 \wedge E_2) \vee \ldots$ is prime, then it is not entailed by its composite consequence $(D_1 \vee D_2 \vee \ldots) \wedge (E_1 \vee E_2 \vee \ldots)$. Only the former requires one's internal state to match the state of the external environment. When the causal explanation depends on the primeness of $(D_1 \wedge E_1) \vee (D_2 \wedge E_2) \wedge \ldots$, as

[14] Proof: $\bigvee_i(D_i \wedge E_i)$ obviously entails $\bigvee_i D_i$. Suppose that $\bigvee_i(D_i \wedge E_i)$ entails a narrow condition D. We must show that $\bigvee_i D_i$ already entails D; we need only show that an arbitrary D_i entails D. Suppose that D_i obtains in a case α. Since E_i is non-trivial, it obtains in some case β. By free recombination, there is a case γ internally like α and externally like β. Since D_i is narrow and E_i is environmental, they both obtain in γ; thus $\bigvee_i(D_i \wedge E_i)$ obtains in γ; since it entails D, D obtains in γ; since D is narrow, it also obtains in α. Thus D_i entails D, as required. Consequently, $\bigvee_i D_i$ is the strongest narrow condition which $\bigvee_i(D_i \wedge E_i)$ entails. By a parallel argument, $\bigvee_i E_i$ is the strongest environmental condition which $\bigvee_i(D_i \wedge E_i)$ entails.

section 3.4 argued that it often will, the extractable narrow condition $D_1 \vee D_2 \vee \ldots$ typically plays no explanatory role; it is a sort of epiphenomenon. What would give the narrow condition an explanatory role is compositeness, not analysability; the arguments for primeness therefore tell against such an explanatory role for the narrow condition.

If an explanation specified an environmental condition E_j, we might combine that with the disjunction $(D_1 \wedge E_1) \vee (D_2 \wedge E_2) \vee \ldots$ to derive the corresponding specific narrow condition D_j (if E_j were incompatible with E_i for every *i* distinct from *j*), which then would play a distinctive explanatory role. But specificity is lack of generality; sections 3.5 and 3.6 showed how lack of generality can be an explanatory vice. An explanation at an appropriate level of generality will be neutral between the disjunct $D_j \wedge E_j$ and some alternative disjunct $D_i \wedge E_i$, while still excluding $D_i \wedge E_j$ and $D_j \wedge E_i$; the unspecific narrow condition $\ldots \vee D_i \vee \ldots \vee D_j \ldots$ extractable from that explanation plays no distinctive explanatory role therein. Non-conjunctive decompositions of the mental into narrow and environmental conditions do not save the internalist picture of the mind, for they do not give narrow conditions the explanatory role which it predicts for them.

4
Anti-Luminosity

4.1 COGNITIVE HOMES

One source of resistance to the conception of knowing as a mental state is the idea that one is guaranteed epistemic access to one's current mental states. According to that idea, one must be in a position to know whether one is in a given mental state, at least when one is attending to the question. When one asks oneself whether one knows a given proposition, one is not always in a position to know the answer. Section 1.2 responded to the objection by arguing that many uncontentious examples of mental states are the same as knowing in this respect. Nevertheless, some are inclined to think that a central core of mental states must be different. If S belongs to that core, then whenever one attends to the question one is in a position to know whether one is in S. In that sense, knowing would not be a core mental state. This chapter argues that there is no central core of mental states in that special sense. That conclusion will be a corollary of a far more general result about the limits of knowledge.

There is a constant temptation in philosophy to postulate a realm of phenomena in which nothing is hidden from us. Descartes thought that one's own mind is such a realm. Wittgenstein enlarged the realm to everything that is of interest to philosophy.[1] That they explained this special feature in very different ways hardly needs to be said; what is remarkable is their agreement on our possession of a cognitive home in which everything lies open to our view. Much of our thinking—for example, in the physical sciences—must operate outside this home, in

[1] See Wittgenstein 1958: §126. Of course, Wittgenstein's 'we', unlike Descartes's, is collective. What is not hidden from each Cartesian subject is only its own thinking, not that of other Cartesian subjects. Wittgenstein is speaking of what is not hidden from any of us. The arguments below have not been, but could be, adjusted to the rejection in Wittgenstein 1969 of claims to know *p* when there is no question of doubting *p*; conversational inappropriateness is compatible with truth. When I feel cold with no question of doubt and I know that everyone else in the room feels cold, I usually know that everyone in the room feels cold. If so, I know that I feel cold.

alien circumstances. The claim is that not all our thinking could be like that.

To deny that something is hidden is not to assert that we are infallible about it. Mistakes are always possible. There is no limit to the conclusions into which we can be lured by fallacious reasoning and wishful thinking, charismatic gurus and cheap paperbacks. The point is that, in our cognitive home, such mistakes are always rectifiable. Similarly, we are not omniscient about our cognitive home. We may not know the answer to a question simply because the question has never occurred to us. Even if something is open to view, we may not have glanced in that direction. Again, the point is that such ignorance is always removable.

The aim of this chapter is to argue that we are cognitively homeless. Although much is in fact accessible to our knowledge, almost nothing is inherently accessible to it. However, it is first necessary to sharpen the issue, to make it more susceptible to argument.

4.2 LUMINOSITY

As in previous chapters, it is convenient to frame the discussion in terms of conditions, which obtain or fail to obtain in various cases. A case depends on a subject (referred to by 'one'), a time (referred to by the present tense), and a possible world. Although conditions are expressed by sentential clauses, they are not propositions as the latter are usually conceived, just because they are open with respect to person, place, and perhaps other circumstances, too. We often use clauses like that, as in 'When it rains, it pours'. The domain of cases will be taken to include counterfactual as well as actual possibilities. Since the cases on which the arguments below rely are physically and psychologically feasible, issues about the bounds of possibility are not pressing.

Conditions are coarsely individuated by the cases in which they obtain: they are identical if they obtain in exactly the same cases. This raises a delicate issue when we say that someone knows that a condition C obtains, for C may be presented in different guises. Under which guise is C known to obtain? If the condition that one is drinking water is the condition that one is drinking H_2O, because they obtain in the same cases, it does not seem to follow that one knows that the condition that one is drinking water obtains if and only if one knows that the condition that one is drinking H_2O obtains, for one may not know that water is H_2O. Fortunately, in a context in which the only relevant presentation of the condition C is as the condition that one is F, knowing that C obtains can be identified with knowing that the condition that one is F obtains, which is in turn only trivially different from knowing that one is F. We can therefore often leave the reference to guises tacit.

We will also use the notion of being in a position to know. To be in a position to know *p*, it is neither necessary to know *p* nor sufficient to be physically and psychologically capable of knowing *p*. No obstacle must block one's path to knowing *p*. If one is in a position to know *p*, and one has done what one is in a position to do to decide whether *p* is true, then one does know *p*. The fact is open to one's view, unhidden, even if one does not yet see it. Thus being in a position to know, like knowing and unlike being physically and psychologically capable of knowing, is factive: if one is in a position to know *p*, then *p* is true. Although the notion of being in a position to know is obviously somewhat vague and context-dependent, it is clear enough for present purposes. The vagueness and context-dependence are in any case primarily the result of fudging in attempts to defend the views to be criticized below.

A condition C is defined to be *luminous* if and only if (L) holds:

(L) For every case α, if in α C obtains, then in α one is in a position to know that C obtains.

Since being in a position to know is factive, the converse of (L) holds for any condition C, so the conditional in (L) could just as well be a biconditional. The picture is that a luminous condition always shines brightly enough to make its presence visible. However, (L) does not say that C must obtain independently of our dispositions to judge that C obtains; for all (L) says, the condition might obtain in virtue of those dispositions.

A realm in which nothing is hidden is a realm in which all conditions are luminous. Our question is: what conditions, if any, are in fact luminous?

Some examples will help. Pain is often conceived as a luminous condition, in the sense that if one is pain, then one is in a position to know that one is in pain (for a recent discussion see McDowell 1989). The definition of luminosity gives scope to finesse some of the more obvious objections to claims of this kind. Thus people who lack the concept of pain—perhaps because their concepts carve up the space of possible sensations in an alternative way—and so never know that they are in pain, may still count as being in a position to know that they are in pain. Perhaps more primitive creatures are sometimes in pain without possessing any concepts at all; if they count as not even being in a position to know that they are in pain, a counterexample to luminosity might still be avoided by a stipulation that the subject of a case must be a possessor of concepts.

Two claims of luminosity are implicit in the following passage from Michael Dummett (1978: 131):

> It is an undeniable feature of the notion of meaning—obscure as that notion is—that meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same.[2]

Thus if two words have the same meaning for one, then one is in a position to know that they have the same meaning; if the words have different meanings for one, then one is in a position to know that they have different meanings. Dummett does not even make the qualification 'in a position to'; what of a subject who has never compared the two words? The two claims of luminosity are genuinely distinct, for the premise that whenever a condition C obtains one is in a position to know that C obtains does not entail the conclusion that whenever C does not obtain one is in a position to know that C does not obtain. If whenever one is awake one is in a position to know that one is awake, it does not follow that whenever one is not awake one is in a position to know that one is not awake (such asymmetries are discussed in Chapter 8). Strictly, of course, having the same meaning and having different meanings are contraries, not contradictories, since both require the words to be meaningful.

Other conditions for which luminosity is often claimed are those of the form: it appears to one that A. When there really is an oasis ahead, one may not be in a position to know that there really is an oasis ahead but, it is supposed, when there at least appears to one to be an oasis ahead, one must be in a position to know that there at least appears to one to be an oasis ahead.

4.3 AN ARGUMENT AGAINST LUMINOSITY

Consider the condition that one feels cold. It appears to have about as good a chance as any non-trivial condition of being luminous. Nevertheless, there is reason to think that it is not really luminous at all. This section presents the argument, and section 4.6 generalizes it. Sections 4.4 and 4.5 discuss objections.

Consider a morning on which one feels freezing cold at dawn, very slowly warms up, and feels hot by noon. One changes from feeling cold

[2] See also Dummett 1981: 632 and 1993: 4. For a recent discussion see Boghossian 1994.

to not feeling cold, and from being in a position to know that one feels cold to not being in a position to know that one feels cold. If the condition that one feels cold is luminous, these changes are exactly simultaneous. Suppose that one's feelings of heat and cold change so slowly during this process that one is not aware of any change in them over one millisecond. Suppose also that throughout the process one thoroughly considers how cold or hot one feels. One's confidence that one feels cold gradually decreases. One's initial answers to the question 'Do you feel cold?' are firmly positive; then hesitations and qualifications creep in, until one gives neutral answers such as 'It's hard to say'; then one begins to dissent, with gradually decreasing hesitations and qualifications; one's final answers are firmly negative.

Let $t_0, t_1, \ldots, t_n$ be a series of times at one millisecond intervals from dawn to noon. Let α_i be the case at t_i ($0 \leq i \leq n$). Consider a time t_i between t_0 and t_n, and suppose that at t_i one knows that one feels cold. Thus one is at least reasonably confident that one feels cold, for otherwise one would not know. Moreover, this confidence must be reliably based, for otherwise one would still not *know* that one feels cold. Now at t_{i+1} one is almost equally confident that one feels cold, by the description of the case. So if one does not feel cold at t_{i+1}, then one's confidence at t_i that one feels cold is not reliably based, for one's almost equal confidence on a similar basis a millisecond earlier that one felt cold was mistaken. In picturesque terms, that large proportion of one's confidence at t_i that one still has at t_{i+1} is misplaced. Even if one's confidence at t_i was just enough to count as belief, while one's confidence at t_{i+1} falls just short of belief, what constituted that belief at t_i was largely misplaced confidence; the belief fell short of knowledge. One's confidence at t_i was reliably based in the way required for knowledge only if one feels cold at t_{i+1}. In the terminology of cases, we have this conditional:

(1_i) If in α_i one knows that one feels cold, then in α_{i+1} one feels cold.

Note that (1_i) is merely a description of a stage in a specific process; it does not purport to be a general principle about feeling cold. Statement (1_i) is asserted for each i from 0 to $n-1$, which is not to say anything about cases other than $\alpha_0, \ldots, \alpha_n$.

Suppose that the condition that one feels cold is luminous. Then in any case in which one feels cold, the condition that one feels cold obtains, so one is in a position to know that the condition that one feels cold obtains, so one is in a position to know that one feels cold; since by hypothesis one is actively considering the matter, one therefore does know that one feels cold. We therefore have this conditional:

(2_i) If in α_i one feels cold, then in α_i one knows that one feels cold.

Now suppose:

(3_i) In α_i one feels cold.

By modus ponens, (2_i) and (3_i) yield this:

(4_i) In α_i one knows that one feels cold.

By modus ponens, (1_i) and (4_i) yield this:

(3_{i+1}) In α_{i+1} one feels cold.

The following is certainly true, for α_0 is at dawn, when one feels freezing cold:

(3_0) In α_0 one feels cold.

By repeating the argument from (3_i) to (3_{i+1}) n times, for ascending values of i from 0 to $n-1$, we reach this from (3_0):

(3_n) In α_n one feels cold.

But (3_n) is certainly false, for α_n is at noon, when one feels hot. Thus the premises (1_0), . . ., (1_{n-1}), (2_0), . . ., (2_{n-1}), and (3_0) entail a false conclusion. Consequently, not all of (1_0), . . ., (1_{n-1}), (2_0), . . ., (2_{n-1}), and (3_0) are true. But it has been argued that (1_0), . . ., (1_{n-1}) and (3_0) are true. Thus not all of (2_0), . . ., (2_{n-1}) are true. By construction of the example, one knows that one feels cold whenever one is in a position to know that one feels cold, so (2_0), . . ., (2_{n-1}) are true if the condition that one feels cold is luminous. Consequently, that condition is not luminous. Feeling cold does not imply being in a position to know that one feels cold.

4.4 RELIABILITY

Since (1_0), . . ., (1_{n-1}) are the key premises in the argument of the last section against luminosity, it is prudent to pause and reconsider the argument for (1_i).

The argument applies reliability considerations to degrees of confidence. These degrees should not be equated with subjective probabilities as measured by one's betting behaviour. For assigning a very high subjective probability to a false proposition does not by itself constitute any degree of unreliability at all, in the sense relevant to knowledge. Suppose that draws of a ball from a bag have been made. The draws are numbered from 0 to 100. You have not been told the results; your information is just that on each draw *i*, the bag contained *i* red balls and 100−*i* black balls. You reasonably assign a subjective probability of *i*/100 to the proposition that draw *i* was red (produced a red ball), and bet accordingly. You know that draw 100 was red, since the bag then contained only red balls, even if the proposition that draw 99 was red—to which you assign a subjective probability of 99/100—is false. That does not justify a charge of unreliability against you. Intuitively, for any *i* less than 100, your bets do not commit you to believing outright that draw *i* was red. Your outright belief may be just that the probability on your evidence that draw *i* was red is *i*/100, which is true. On draw 100, unlike the others, you can form the belief on non-probabilistic grounds that it was red. What incurs the charge of unreliability is believing a false proposition outright, not assigning it a high subjective probability.

What is the difference between believing *p* outright and assigning *p* a high subjective probability? Intuitively, one believes *p* outright when one is willing to use *p* as a premise in practical reasoning. Thus one may assign *p* a high subjective probability without believing *p* outright, if the corresponding premise in one's practical reasoning is just that *p* is highly probable on one's evidence, not *p* itself. Outright belief still comes in degrees, for one may be willing to use *p* as a premise in practical reasoning only when the stakes are sufficiently low. Nevertheless, one's degree of outright belief in *p* is not in general to be equated with one's subjective probability for *p*; one's subjective probability can vary while one's degree of outright belief remains zero. Since using *p* as a premise in practical reasoning is relying on *p*, we can think of one's degree of outright belief in *p* as the degree to which one relies on *p*. Outright belief in a false proposition makes for unreliability because it is reliance on a falsehood. The degrees of confidence mentioned in the argument for (1_i) should therefore be understood as degrees of outright belief.

The argument for (1_i) assumes that the underlying basis on which one believes that one feels cold changes at most slightly between t_i and t_{i+1}, for otherwise an error in the belief at t_{i+1} might not threaten the reliability of the belief at t_i. For example, if one believes inferentially at t_{i+1} and not at all inferentially at t_i, false belief at t_{i+1} might well be consistent with knowledge at t_i. Apparent gradualness in the process does not guarantee gradualness at the underlying level (Wright 1996: 937). Nevertheless, we can choose an example in which there is gradualness at the underlying level too, and that will suffice for a counterexample to (L). The basis on which one judges that one feels cold need not change suddenly as one gradually becomes colder.

The invocation of reliability does not presuppose that whether one feels cold is independent of one's dispositions to judge that one does. Luminosity is often supposed to rest on a constitutive connection between the obtaining of the condition and one's judging it to obtain, but the effect of such a connection would be to make reliability less contingent, not to make unreliability consistent with knowledge.

The concept of reliability is notoriously vague. If one believes *p* truly in a case α, in which other cases must one avoid false belief in order to count as reliable enough to know *p* in α? There is no obvious way to specify in independent terms which other cases are relevant. This is sometimes known as the *generality problem* for reliabilism. Some have argued that the generality problem is insoluble and that reliabilist theories in epistemology should therefore be abandoned (Conee and Feldman 1998). Let us concede for the sake of argument that the generality problem is indeed insoluble. It does not follow that appeals to reliability in epistemology should be abandoned. For the insolubility of the generality problem means that the concept of reliability cannot be defined in independent terms; it does not mean that the concept is incoherent. Most words express indefinable concepts; 'reliable' is not special in that respect. Irrespective of any relation to the concept *knows*, we clearly do have a workable concept *is reliable*; for example, historians sensibly ask which of their sources are reliable. The concept is certainly vague, but most words express vague concepts; 'reliable' is not special in that respect either. The concept *is reliable* need not be precise to be related to the concept *knows*; it need only be vague in ways that correspond to the vagueness in *knows*. No reason has emerged to doubt the intuitive claim that reliability is necessary for knowledge.

If one believes *p* truly in a case α, one must avoid false belief in other cases sufficiently similar to α in order to count as reliable enough to know *p* in α. The vagueness in 'sufficiently similar' matches the vagueness in 'reliable', and in 'know'. Since the account of knowledge developed in Chapter 1 implies that the reliability condition will not be a conjunct in a non-circular analysis of the concept *knows*, we need not even assume that we can specify the relevant degree and kind of similarity without using the concept *knows*. To suppose that reliability is necessary for knowledge is not to suppose that the concept *knows* can be analysed in terms of the concept *is reliable*, for it may be impossible to frame other necessary conditions without use of the concept *knows* whose conjunction with reliability is a necessary and sufficient condition for knowledge (see section 1.3).

We cannot expect always to apply a vague concept by appeal to rigorous rules. We need good judgement of particular cases. Indeed, even when we can appeal to rigorous rules, they only postpone the moment at which we must apply concepts in particular cases on the basis of good judgement. We cannot put it off indefinitely, on pain of never getting started. The argument for (1_i) appeals to such judgement. The intuitive idea is that if one believes outright to some degree that a condition C obtains, when in fact it does, and at a very slightly later time one believes outright on a very similar basis to a very slightly lower degree that C obtains, when in fact it does not, then one's earlier belief is not reliable enough to constitute knowledge. The earlier case is sufficiently similar to the later case. One's earlier reliance on C has too much in common with one's later reliance on it. The use of the concept *is reliable* here is a way of drawing attention to an aspect of the case relevant to the application of the concept *knows*, just as one might use the concept *is reliable* in arguing that a machine ill serves its purpose. The aim is not to establish a universal generalization but to construct a counterexample to one, the luminosity principle (L). As with counterexamples to proposed analyses of concepts, we are not required to derive our judgement as to whether the concept applies in a particular case from general principles.

Within the limits just explained, we can nevertheless see how a reliability condition on knowledge is consonant with the role of knowledge in the causal explanation of action, as described in sections 2.4 and 3.4. Knowledge is superior to mere true belief because, being more robust in the face of new evidence, it better facilitates action at a temporal distance. Other things being equal, given rational sensitivity to new evidence, present knowledge makes future true belief more likely than mere present true belief does. This is especially clear when the future belief is in a different proposition, that is, when the future belief can differ in truth-value from the present belief.

Some hunters see a deer disappear behind a rock. They believe truly that it is behind the rock. To complete their kill, they must maintain a true belief about the location of the deer for several minutes. But since it is logically possible for the deer to be behind the rock at one moment and not at another, their present-tensed belief may be true at one moment and false at another. By standard criteria of individuation, a proposition cannot change its truth-value; the sentence 'The deer is behind the rock' expresses different propositions at different times. In present terminology, it is logically possible for the unchanging condition that the deer is behind the rock to obtain at one moment and not at another. If the hunters know that the deer is behind the rock, they have the kind of sensitivity to its location that makes them more likely to have future true beliefs about its location than they are if they merely believe truly that it is behind the rock. If we are to explain why they later succeeded in killing the deer, given the foregoing situation, then it is more relevant that they know that the deer is behind the rock than that they believe truly that it is behind the rock.

The role of knowledge in the explanation of action exploits a kind of reliability. If at time t on basis b one knows p, and at a time t^* close enough to t on a basis b^* close enough to b one believes a proposition p^* close enough to p, then p^* should be true. The argument of section 4.3 allows us to pick t^*, b^*, and p^* arbitrarily close to t, b, and p respectively. We can make the time interval between t_i and t_{i+1} as short as we like. Since the relevant beliefs are in the obtaining of the same condition at those times, they will be correspondingly close. Since, as noted above, the beliefs can also be assumed to change in basis only gradually, their bases too will be correspondingly close. A well-chosen example will verify (1_0), . . ., (1_{n-1}) and thereby provide the required counterexample to (L). A reliability condition on knowledge facilitates the role that knowledge does in fact play in the causal explanation of action. The appeal to such a condition does not depend only on brute intuition; it fits the independently motivated conception of knowing as a mental state.

4.5 SORITES ARGUMENTS

An obvious doubt arises about the argument of section 4.3. The reasoning is very reminiscent of that in sorites paradoxes. If with 0 hairs on one's head one is bald, and, for every natural number i, with i hairs on one's head one is bald only if with $i + 1$ hairs on one's head one is bald, then for any natural number n, however large, it follows that with n hairs on one's head one is bald. The reasoning may therefore be suspected of concealing a mistake just like the concealed mistake in sorites reasoning, whatever that is. Does the argument illicitly exploit the vagueness of 'feels cold' or 'know'?

The doubt can be made more specific. If the conclusion of the argument is false, then either not all the premises are true or the reasoning is invalid. Given (1_0), . . ., (1_{n-1}) and the straightforwardly true (3_0) as auxiliary premises, the argument derives (2_0), . . ., (2_{n-1}) from the supposed luminosity of the condition at issue and uncontested background assumptions, and then uses modus ponens to reach the straightforwardly false (3_n). By reductio ad absurdum, luminosity is rejected. On any reasonable view of vagueness, this reasoning shows that the luminosity claim is less than perfectly true, given that (1_0), . . ., (1_{n-1}) are perfectly true.

On some accounts, the rule of modus ponens fails to preserve less than perfect truth, because it sometimes leads from almost perfectly true premises to a conclusion that is not even almost perfectly true. But modus ponens should still preserve perfect truth. Within degree-theoretic semantics, a pseudo-conditional can be defined for which a conditional statement is perfectly true if and only if its consequent is at worst slightly less true than its antecedent (Peacocke 1981: 127). For present purposes, however, we can legitimately stipulate that the conditional to be used in the argument is of the more conventional kind for which the conditional statement is perfectly true if and only if the consequent is at least as true as the antecedent.

On other accounts, the rule of reductio ad absurdum is problematic because an assumption can have perfectly false consequences without itself being perfectly false, and therefore without having a perfectly true negation. Nevertheless, an assumption with perfectly false consequences is still less than perfectly true. Moreover, it is arguable that vagueness requires no revision of classical logic at all.[3]

For the purposes of this chapter, it would suffice to argue that the luminosity claim is less than perfectly true, for then it will have perfectly false consequences, which should discourage its application to philosophy. Thus the way for the defender of (perfect) luminosity to use the connection with sorites paradoxes is by arguing that not all of (1_0), . . ., (1_{n-1}) are perfectly true, and using the vagueness of some relevant term to explain away their plausibility. Of course, the argument for (1_i) would remain to be addressed. Fortunately, however, the strategy can be tested more directly. For if (1_0), . . ., (1_{n-1}) are in effect the premises of a sorites paradox, then sharpening the relevantly vague expressions should make at least one of them clearly false, just as sharpening the term 'bald' by stipulating a cut-off point gives the conditional 'With *i* hairs on one's head one is bald only if with *i* + 1 hairs on one's head one is bald' a clearly false instance. Does the same happen here?

The relevantly vague expressions in (1_i) are 'feels cold' and 'knows'. We can sharpen 'feels cold' by using a physiological condition to resolve borderline cases. Let us assume that the subject of the process has no access to the technology needed to determine whether the physiological condition obtains, and so is not in a position to know whether it does. These stipulations in no way weaken the argument for (1_i). The considerations about reliability remain as cogent as before, for they were

[3] See Williamson 1994b on logic for vague languages. The arguments of the present chapter do not depend on the epistemic account of vagueness developed there, which requires no revision of classical logic.

based on our limited powers of discrimination amongst our own sensations, not on the vagueness of 'feels cold'. It might be objected that the sharpening violates the intended meaning of 'feels cold'. However, that would not undermine the contrast between (1_i) and the major premise of a sorites paradox. For *any* complete sharpening of 'bald' yields a clearly false instance of the principle 'With i hairs on one's head one is bald only if with $i + 1$ hairs on one's head one is bald', even if it violates the intended meaning of 'bald' by, for example, falsifying the converse downwards principle 'With $i + 1$ hairs on one's head one is bald only if with i hairs on one's head one is bald'. By definition, the sharpened term applies wherever the unsharpened term clearly applied and fails to apply wherever the unsharpened term clearly failed to apply; thus, on any sharpening, 'With 0 hairs on one's head one is bald' is true and 'With i hairs on one's head one is bald' is false for a suitably large number n, so for some number i the conditional 'With i hairs on one's head one is bald only if with $i + 1$ hairs on one's head one is bald' is false. Thus even the truth of (1_0), . . ., (1_{n-1}) on a sharpening of the vague terms that violates their intended meaning is enough to differentiate them from the premises of a sorites paradox.

The vague expression 'knows' remains. Sharpen it by tightening up its conditions of application: in the new sense it is not to apply in borderline cases for knowing in the old sense. It does not matter whether it applies in borderline cases of borderline cases for the old sense. If anything, this strengthens the argument for (1_i), by building more into its antecedent. It does not help one to know whether one feels cold. Indeed, one need not even be aware of the stipulation about 'know', for it is made by the theorist, not by the subject.

The stipulations will not make 'feels cold' and 'knows' perfectly precise; no feasible sharpening could do that. Fortunately, perfect precision is not necessary. We need only sharpen those expressions enough to resolve the finitely many borderline cases that actually arise in the argument. Such sharpening has the opposite effect to that predicted by the assimilation of the argument against luminosity to sorites reasoning; (1_i) becomes more not less plausible. The argument is not just another sorites paradox.

Nevertheless, the argument against luminosity might be thought to commit a subtler fallacy of vagueness. A defender of (2_i) might take the vagueness of its constituent terms to be essential to its truth, and explain the plausibility of (1_i) by assigning it a status short of perfect truth, while conceding that all of (1_0), . . ., (1_{n-1}) are true on some sharpenings, such as those considered above. The critic might take any sharpening that falsifies (2_i) to violate the intended meanings of the vague terms, on the grounds that those meanings make (2_i) analytic. On such a view, some unsharpened (1_i) would be almost but not quite perfectly true, because its consequent would be almost but not quite as true as its antecedent. The reliability conditions adduced in favour of (1_i) would be treated as almost but not quite perfectly correct. No justification has been provided for not treating them as perfectly correct, but let that pass. For the concession is in any case inadequate. The defender of (2_i) must reject the following variation on (1_i):

($1P_i$) If it is perfectly true that in α_i one knows that one feels cold, then it is perfectly true that in α_{i+1} one feels cold.

For if (2_i) is perfectly true, then the perfect truth of its antecedent implies the perfect truth of its consequent:

($2P_i$) If it is perfectly true that in α_i one feels cold, then it is perfectly true that in α_i one knows that one feels cold.

Statements ($1P_i$) and ($2P_i$) give an argument from the perfect truth of (3_i) to the perfect truth of (3_{i+1}), and therefore from the uncontested perfect truth of (3_0) to the perfect truth of (3_n); but the falsity of (3_n) is uncontested.

The critic will presumably treat ($1P_i$) like (1_i), claiming that for some number *i*, it can be perfectly true that in α_i one knows that one feels cold, but slightly less than perfectly true that in α_{i+1} one feels cold. Can there be such an *i*? If it is less than perfectly true that in α_{i+1} one feels cold, then there is a strict standard by which it is false in α_{i+1} that one feels cold; so, by that standard, in α_{i+1} one is fairly confident of what is false, that one feels cold. If so, it is less than perfectly true that in α_i one knows that one feels cold, if the reliability considerations are to be assigned any positive weight at all. To put the argument more directly, if it is perfectly true that in α_i one knows that one feels cold, then it is perfectly true that one achieves the level of reliability necessary for knowing, and therefore perfectly true that in α_{i+1} one feels cold. Thus the objection to ($1P_i$) fails, and ($1P_0$), . . ., ($1P_{n+1}$) suffice for an argument that not all of (2_0), . . ., (2_{n-1}) are perfectly true. Invoking degrees of truth will not protect claims of perfect luminosity.

The point is reinforced by the observation that, once the luminosity assumption is dropped, (3_n) does not follow in classical logic from (1_0), . . ., (1_{n-1}) and (3_0). To see this, pick *j* and *k* such that $0 \leq j < k < n$; for each *i*, evaluate 'One feels cold' as true in α_i if and only if $i \leq k$, and otherwise as false; evaluate 'One knows that one feels cold' as true in α_i if and only if $i \leq j$, and otherwise as false. On this evaluation, (1_i) is always true, for if the antecedent is true, then $i \leq j < k$, so $i + 1 \leq k$, so the consequent is true. Statement (3_0) is true because $0 < k$. Statement (3_n) is false because $k < n$. We can extend this evaluation in the manner of the standard semantics for modal logic by treating cases like possible worlds and 'One knows that . . .' like 'It is necessary that . . .'. The foregoing evaluation results if one defines a case α_h to be accessible from a case α_i if and only if $|h-i| \leq k-j$, evaluates 'One knows that A' as true at a case α_i if and only if 'A' is true at all cases accessible from α_i, and evaluates 'one feels cold' as before. Since a classical evaluation makes (1_0), . . ., (1_{n-1}) and (3_0) true and (3_n) false, the latter does not follow from the former in classical logic. Contrast the sorites paradox: for any n, 'With n hairs on one's head one is bald' does follow in classical logic from 'With 0 hairs on one's head one is bald' and conditionals of the form 'With i hairs on one's head one is bald only if with $i + 1$ hairs on one's head one is bald'. Once luminosity is denied, conditionals of the form (1_i) generate no paradox.

Consistently with all this, we can postulate a more general phenomenon of which both vagueness and failures of luminosity independent of vagueness are special cases (Williamson 1994b and below). On such a view, the epistemological principles underlying (1_i) are important for vagueness too, but it does not follow that all their manifestations involve vagueness. Indeed, the epistemological principles by themselves imply no specific theory of vagueness.

4.6 GENERALIZATIONS

Section 4.3 argued that a specimen condition—that one feels cold—is not luminous. How far does the argument generalize?

The argument assumed nothing specific about the condition of feeling cold. It extends to the examples of supposedly luminous conditions mentioned in section 4.2. Since pain sometimes gradually subsides, for example, an argument against the luminosity of the condition that one is in pain can be modelled on the argument against the luminosity of the condition that one feels cold, without any structural revisions. It is not perfectly true that whenever one is in pain, one is in a position to know that one is in pain. That one is in pain does not imply that one is in a position to know that one is in pain. Similarly, two synonyms can gradually diverge in meaning, as a mere difference in tone grows into a difference in application. The structure of the argument against luminosity is just as before. That two words have the same meaning for one does not imply that one is in a position to know that they have the same meaning for one. Equally, that they have different meanings for one does not imply that one is in a position to know that they have different meanings for one. The argument also applies to the condition that things appear to one in some way, for example, that it looks to one as though there is a purple patch ahead. Cases in which things appear to one in some way can gradually give way to cases in which they do not appear to one in that way. That they appear to one in that way does not imply that one is in a position to know that they appear to one in that way.

The condition that things appear to one in some way is often supposed to be a paradigm of what is called *response-dependence*. Unfortunately, that phrase is used in many senses, few of them clear. If the response-dependence of a condition means only that whether it obtains has *some* constitutive dependence on whether one is disposed to judge that it obtains, then response-dependence does not entail luminosity, although non-luminosity does constrain what forms of dependence a condition can exhibit (see Williamson 1994b: 180–4 for the case of colour). But if 'response-dependent' is so defined that a response-dependent condition must be luminous, then the conditions that are standardly taken as paradigms of response-dependence are none of them response-dependent.

Further applications of the argument involve conditions on one's knowledge. Since one can gain or lose knowledge gradually, we can use the argument to show that, for most propositions *p*, neither the condition that one knows *p* nor the condition that one does not know *p* is luminous. One can know *p* without being in a position to know that one knows *p*, and one can fail to know *p* without being in a position to know that one fails to know *p*. Chapters 5 and 8 respectively discuss these applications in more detail.

On what general features of a condition does the argument against luminosity depend? As it stands, it requires the condition to obtain in some cases and not in others. Thus it is ineffective against a condition that obtains in all cases or in none. Given a sufficiently restrictive understanding of what a case is, that might include the Cartesian condition that one exists, or even that one thinks. It does not include the condition that one is thinking about one's existence, for one does that in some cases and not in others on any reasonable understanding of what a case is.

A condition that obtains in no case, the impossible condition, is automatically luminous; (L) holds vacuously. Is a condition that obtains in every case, the necessary condition, luminous too? It is luminous as presented in a simple tautological guise, if cases are restricted to those in which the subject has the concepts to formulate the tautology. It is not luminous as presented in the guise of an a posteriori necessity, or an unproved mathematical truth, or if the cases include some in which one lacks appropriate concepts.

The argument also requires the possibility of a change from cases in which the condition obtains to cases in which it does not. Thus it would not be effective against an eternal condition, which always obtains if it ever obtains: for example, the condition that one felt cold at midnight on New Year's Eve 1999. However, many eternal conditions, including that one, permit a change from cases in which one is in a position to know that they obtain to cases in which one is not in a position to know that they obtain. Such a condition cannot be luminous, for since it obtains in the earlier cases in which one is in a position to know that it obtains (because being in a position to know is factive), it also obtains in the later cases in which one is not in a position to know that it obtains (because the condition is eternal). Thus an eternal condition is luminous only if one cannot change from being in a position to know that it obtains to not being in such a position. There are candidates for such conditions. For example, if a subject S is always in a position to know that she is S—which is not to say that she must know her own name—then anyone who is ever in a position to know that the condition that one is S obtains is always in a position to know that it obtains, because the only such person is S herself. Perhaps the argument could be extended to show that not even this condition is luminous, by consideration of a science-fiction process in which someone else is gradually replaced by S. However, no such extension will be attempted here. Such examples do not seriously threaten the idea that only trivial conditions are luminous.

The argument also assumes that one is considering the relevant condition under the relevant guise throughout the process. Consequently, it does not apply to some conditions on one's considerations. For example, let C be the condition that one is entertaining the proposition that it is raining, and let G be the guise under which C has just been presented here. To consider C under G is to consider as such the condition that one is entertaining the proposition that it is raining; in so doing, one thereby entertains the proposition that it is raining, so C obtains. Thus one cannot gradually pass from cases in which C obtains to cases in which C does not obtain while considering C under G throughout the process. Although one can gradually pass from cases in which C obtains to cases in which C does not obtain, one does not consider C under G in the late stages of the process. For all the argument shows, C is luminous: if one is entertaining the proposition that it is raining, then one is in a position to know that one is entertaining the proposition that it is raining. When one is entertaining a slightly different proposition *p*, one does not have a high degree of false belief that one is entertaining the proposition that it is raining; one has a high degree of true belief that one is entertaining *p*, since the belief derives its content from *p* itself. Thus the argument does not apply to examples in which one considers the condition only when it obtains. Such examples constitute a very minor limitation on the generality of the argument. In any case, we may conjecture that, for any condition C, if one can move gradually to cases in which C obtains from cases in which C does not obtain, while considering C throughout, then C is not luminous. The conjecture is discussed further in section 5.2.

Luminous conditions are curiosities. Far from forming a cognitive home, they are remote from our ordinary interests. The conditions with which we engage in our everyday life are, from the start, non-luminous.

4.7 SCIENTIFIC TESTS

To be physically and psychologically capable of knowing *p* is not sufficient, even given *p*, for being in a position to know *p*; one may be in the wrong place. Thus it does not follow from the non-luminosity of a condition that there are cases in which, although it obtains, one is not physically and psychologically capable of knowing that it obtains. Nevertheless, it is natural to ask, if one is not in a position to know in a case α that one then feels cold, how is one to know in some other case β that in α one feels cold? Must or can there be such a case β? Analogous questions arise about other non-luminous conditions. The argument of section 4.3 leaves them open. It is consistent with, but does not entail, the possibility of a physiological technique by which one could subsequently discover that one had been feeling cold in α.

The hypothetical technique faces difficulties. Suppose that feelings of cold and hot are found generally to be correlated with a measurable physiological variable V. We must discover which values of V are associated with the condition that one feels cold. They include the values associated with the condition that one is in a position to know that one feels cold. But if they included only those values, the condition that one feels cold would be luminous, which it is not. We are not in a position to know which further values of V are associated with that condition. Our problem is that we cannot calibrate the physiological measurement of feeling cold. Even if measurements of V were perfectly precise—which they will not be—they would not answer the original question. Attempts to measure other ordinary conditions face similar problems. Their non-luminosity prevents us from perfectly calibrating instruments to detect whether they obtain.

It might still be held to be metaphysically possible to find out whether one feels cold by the testimony of a literal or metaphorical *deus ex machina*. But it certainly cannot be assumed without argument that if an ordinary condition obtains in a case α, then in some possible case it is known that the condition obtains in α. Section 12.5 discusses the issue further.

4.8 ASSERTIBILITY CONDITIONS

The failure of luminosity impinges on Michael Dummett's arguments for an anti-realist theory of meaning, which explains meanings in terms of the conditions under which speakers are warranted in using sentences assertively, by contrast with a realist theory of meaning, which explains meanings in terms of the conditions under which sentences express truths. Of course, Dummett's anti-realist does not make the extreme claim that every condition is luminous. All parties can accept that stone age men lived when the moon caused the tides, although they were not in a position to know that the moon caused the tides. The connection between luminosity and anti-realism is a subtler one.

Dummett objects to the realist's truth-conditional theory of meaning that it violates a necessary connection between meaning and use. To understand a sentence is to know what it means. If, as Dummett's realist holds, meanings are truth-conditions, then speakers of a language know the truth-conditions of its sentences.[4] Knowing the truth-condition of a sentence *s* cannot consist merely in being disposed to say something of the form '*s* is true if and only if P'; one must also understand the biconditional, and an infinite regress looms. In the basic case, one's knowledge of the truth-condition must be implicit. If one could always

[4] Provision must naturally be made for non-declarative sentences, perhaps by a distinction between sense and force. Even for declaratives, knowledge of truth-conditions may not suffice for understanding. More generally, knowledge of meaning may not suffice for understanding. A reliable informant tells me that a sentence of Hungarian written on the board means that the cat sat on the mat. I know by testimony that the sentence means that the cat sat on the mat, but arguably I do not understand it because I do not know what each constituent word contributes to that meaning. Such complications will be ignored in what follows.

recognize whether it obtained, then knowledge of the truth-condition of *s* might consist in a willingness to assert *s* just when the truth-condition obtained. Dishonesty, shyness, and other complications are assumed to have been somehow filtered out. However, the realist insists that the truth-conditions of some sentences obtain even though no speaker of the language can recognize that they obtain. Dummett argues that the realist has no substantial explanation of what knowing that sentences have those truth-conditions consists in. The proposed remedy is that the meaning of a sentence should be given by its assertibility-condition rather than by its truth-condition. Thus to understand a sentence is to know its assertibility-condition, and this knowledge can consist in a willingness to assert the sentence just when the assertibility-condition obtains.[5]

The remedy fails if the objection to truth-conditional theories of meaning applies equally to assertibility-conditional theories of meaning. Thus Dummett's argument requires that when an assertibility-condition obtains, competent speakers of the language can recognize that it obtains. He acknowledges that requirement: 'The conditions under which a sentence is recognized as true or false . . . have, by the nature of the case, to be conditions which we can recognize as obtaining when they obtain' (1981: 586; compare 1991: 317 and 1993: 45–6). That is, when a recognition-condition obtains, we can recognize that the recognition-condition obtains. Dummett evidently intends the recognition-condition for the truth of a sentence to be its assertibility-condition, which yields the thesis that when an assertibility-condition obtains, we can recognize that it obtains. But recognizing is coming to know, and Dummett's 'can' may be glossed as 'is in a position to'. Thus Dummett requires assertibility-conditions to be luminous.

The argument against luminosity in section 4.3 generalizes to assertibility-conditions. For example, it can gradually cease to be assertible that it is raining. By the argument, that it is assertible that it is raining does not imply that one is in a position to know that it is assertible that it is raining. Even in the mathematical case, in which Dummett uses the proof-based intuitionistic semantics as a paradigm of an assertibility-conditional theory of meaning, proofs can be understood or forgotten gradually.[6] By the argument, that one has a proof of a mathematical assertion does not imply that one is in a position to know that one has a

[5] Dummett says that 'no difficulty can any longer arise over what such knowledge consists in' (1977: 375). Note that assertibility-conditions are not truth-conditions even on Dummett's anti-realist conception of truth (see n. 7 below).

[6] On recognizing intuitionistic proofs as proofs see Weinstein 1983 and Pagin 1994.

proof of it. Thus assertibility-conditions have the very feature that is supposed to lay truth-conditions open to Dummett's attack.[7]

An assertibility-conditional theory of meaning is likely to distinguish between canonical and non-canonical warrants for assertion, for example, between having a proof and having been told by a reliable informant that there is one. The recursive semantics will be formulated in terms of canonical warrants; a non-canonical warrant will be explained as an entitlement to believe that there is a canonical warrant. The argument applies whether or not 'warrant' is qualified by 'canonical'.

The anti-realist might reply that one's understanding of a sentence can consist in the fact that one is willing to assert it when and only when its assertibility-condition obtains, even if one does not know that it obtains. This reply concedes that assertibility-conditions fail Dummett's luminosity constraint; but then something is wrong with his argument for assertibility-conditional theories of meaning, which treats that constraint as binding.

A different reply is that if Dummett intends recognizability to be assertibility, then what he requires is only that when *p* is assertible, it is assertible that *p* is assertible. If misleading evidence sometimes warrants false assertions, then it might be assertible that *p* is assertible even when one is not in a position to know that *p* is assertible, so Dummett would not require assertibility-conditions to be luminous. This reply fails because the argument of section 4.3 can be generalized to an argument that no non-trivial condition obtains only when it is assertible that it obtains.[8]

[7] In criticizing truth-conditional theories of meaning, Dummett often focusses on the *undecidability* of truth, that is, on speakers' lack of an effective procedure for coming to know whether a given truth-condition obtains. When the argument takes this form, it threatens assertibility-conditional theories of meaning too unless assertibility is decidable, that is, unless speakers always have an effective procedure for coming to know whether a given assertibility-condition obtains (Crispin Wright, 1992b: 56, endorses the decidability of assertibility). The decidability of assertibility does not entail the decidability of truth, even on most anti-realist conceptions of truth, for the latter do not identify truth with assertibility in non-ideal cases. For example, intuitionists identify assertibility with possession of a proof, and truth with the existence of one (that is, with the possibility of possessing it); they deny the law of excluded middle just because they take truth to be undecidable. Even if assertibility were luminous, its decidability would follow only on the additional assumption that unassertibility is luminous too. Thus what the anti-realist requires of assertibility depends on whether what realist truth is accused of crucially lacking is luminosity or decidability, even though, given bivalence and the presence of classical negation in the object language—both assumptions acceptable to the realist—truth is decidable if and only if it is luminous (on an appropriate construal of 'in a position to know'). But since any decidable condition is luminous (on that construal), the points in the text apply to both forms of anti-realist argument.

[8] A conditional of the form 'If in α_i it is assertible that C obtains, then in α_{i+1} C obtains' replaces (1_i) (Williamson 1995a). Chapter 11 defends a view on which the move

Dummett presents his argument as a challenge to the realist to explain what knowledge of realist truth-conditions consists in. He does not claim to prove that the realist cannot meet the challenge, although he denies that it has been met so far. He allows that it might be met in some areas and not in others. But he assumes that the anti-realist can easily meet the corresponding challenge, to explain what knowledge of assertibility-conditions consists in. If the foregoing argument is correct, that assumption is false; the anti-realist faces the same sort of difficulty as the realist does. The contrast between truth-conditions and assertibility-conditions is off the point.

Both truth-conditional and assertibility-conditional theories of meaning find it hard to meet Dummett's challenge because both truth-conditions and assertibility-conditions are non-luminous. They share this feature with every other kind of non-trivial condition that might be offered as the meaning of a sentence. Since trivial conditions are not serious candidates for the meanings of most sentences, a serious X-conditional theory of meaning will find it hard to meet Dummett's challenge, for any X. If any systematic theory of meaning can be cast as X-conditional for some X, then any systematic theory of meaning will find it hard to meet Dummett's challenge. If 'hard' turns out to be 'impossible', then failure to meet the challenge eliminates truth-conditional theories of meaning only if it eliminates all systematic theories of meaning. The challenge embodies extreme demands on a theory of meaning. We should not assume the possibility of a reductive explanation of what knowledge of meaning 'consists in' of the kind that Dummett demands.

On an anti-realist picture, thought initially engages with conditions whose *esse* is their *percipi*; if it later finds its laborious way to conditions of greater depth, it must do so from the starting point of that cognitive home. Assertibility-conditions are pictured as forming a cognitive home in language. They do not. Thought engages with conditions whose *esse* is distinct from their *percipi* as soon as it engages with any conditions at all; even perception does. Trivialities aside, there is nothing else to engage with. We have no cognitive home.

in question makes no difference, because only knowing *p* warrants asserting *p*. Crispin Wright 1992b: 18 seems to assume the S4 principle for assertibility in arguing that truth and assertibility coincide in 'positive normative force'.

5
Margins and Iterations

5.1 KNOWING THAT ONE KNOWS

One can know something without being in a position to know that one knows it. We reached that conclusion using the form of argument developed in the previous chapter, for by a gradual process one can gain or lose knowledge. Similarly, one can know that one knows something without being in a position to know that one knows that one knows it, for by a gradual process one can gain or lose knowledge that one knows. This chapter explores such limits to our ability to iterate knowledge. They stem from our need of margins for error in much of our knowledge. Those limits make problems for common knowledge, in which everyone knows that everyone knows that everyone knows that . . .. Chapter 6 will apply the results to suggest a diagnosis of the paradox of the Surprise Examination and related puzzles.

We first consider in some detail a variant argument against the luminosity of the condition that one knows something. One can know without being in a position to know that one knows.

Looking out of his window, Mr Magoo can see a tree some distance off. He wonders how tall it is. Evidently, he cannot tell to the nearest inch just by looking. His eyesight and ability to judge heights are nothing like that good. Since he has no other source of relevant information at the time, he does not know how tall the tree is to the nearest inch. For no natural number *i* does he know that the tree is *i* inches tall, that is, more than *i*−0.5 and not more than *i*+0.5 inches tall. Nevertheless, by looking he has gained some knowledge. He knows that the tree is not 60 or 6,000 inches tall. In fact, the tree is 666 inches tall, but he does not know that. For all he knows, it is 665 or 667 inches tall. For many natural numbers *i*, he does not know that the tree is not *i* inches tall. More precisely, for many natural numbers *i*, he does not know the proposition expressed by the result of replacing '*i*' in 'The tree is not *i* inches tall' by a numeral designating *i*. We are not concerned with knowledge of propositions expressed by sentences in which *i* is designated by a definite description, such as 'the height of the tree in inches', for he may not know which number fits the description.

To know that the tree is *i* inches tall, Mr Magoo would have to judge that it is *i* inches tall; but even if he so judges and in fact it is *i* inches tall, he is merely guessing; for all he knows it is really $i-1$ or $i+1$ inches tall. He does not know that it is not. Equally, if the tree is $i-1$ or $i+1$ inches tall, he does not know that it is not *i* inches tall. Anyone who can tell by looking that the tree is not *i* inches tall, when in fact it is $i+1$ inches tall, has much better eyesight and a much greater ability to judge heights than Mr Magoo has. These reflections do not depend on the value of *i*. For *no* natural number *i* is the tree $i+1$ inches tall while he knows that it is not *i* inches tall. In this story, Mr Magoo reflects on the limitations of his eyesight and ability to judge heights. Mr Magoo knows the facts just stated. Consequently, for each relevant natural number *i*:

(1_i) Mr Magoo knows that if the tree is $i+1$ inches tall, then he does not know that the tree is not *i* inches tall.

We could make the case for (1_i) even stronger by reducing the interval of an inch to something much smaller, perhaps a millionth of an inch, but that should not be necessary. To make the conditional 'If the tree is $i+1$ inches tall, then he does not know that it is not *i* inches tall' as uncontentious as possible, we can read 'if' as the truth-functional conditional, the weakest of all conditionals. In effect, it merely denies the conjunction 'The tree is $i+1$ inches tall and he knows that it is not *i* inches tall'.

Suppose, for a reductio ad absurdum, that the condition that one knows a proposition is luminous: if one knows it, then one is in a position to know that one knows it. We may also assume that, in the case at hand, for each proposition *p* pertinent to the argument, Mr Magoo has considered whether he knows *p*. Consequently, if he is in a position to know that he knows *p*, he does know that he knows *p*. Thus:

(KK) For any pertinent proposition *p*, if Mr Magoo knows *p* then he knows that he knows *p*.

Statement (KK) is a special case of the general 'KK' principle that if one knows something then one knows that one knows it, but sufficiently restricted to avoid many of the objections to the latter (for some of which see Sorensen 1988: 242). For example, (KK) does not imply by iteration that if *p* is pertinent then Mr Magoo has every finite number of iterations of knowledge of *p*, for it has not been granted that if *p* is pertinent then so too is the proposition that he knows *p*. The pertinent propositions are just those that occur in the argument below, which form a strictly limited set. Statement (KK) is also immune to the objection that a simple creature without the concept *knows* might still know, but would not know that it knew, for Mr Magoo has the concept *knows*.

We may legitimately assume that in the example Mr Magoo has been reflecting on the height of the tree and his knowledge of it so carefully that he has drawn all the pertinent conclusions about its height that follow deductively from what he knows; he has thereby come to know those conclusions. Let us consider a time at which that process is complete. We can therefore assume:

(C) If p and all members of the set X are pertinent propositions, p is a logical consequence of X, and Mr Magoo knows each member of X, then he knows p.

Of course, (C) is not justified by some general closure principle about knowledge. We often fail to know consequences of what we know, because we do not know that they are consequences. Statement (C) is simply a description of Mr Magoo's state once he has attained reflective equilibrium over the propositions at issue, by completing his deductions. Since Mr Magoo's deductive capacities do not fully enable him to overcome the limitations of his eyesight and ability to judge heights, and he knows that they do not, (1_i) remains true for all i.

By (KK), we can infer (3_i) from (2_i):

(2_i) Mr Magoo knows that the tree is not i inches tall.

(3_i) Mr Magoo knows that he knows that the tree is not i inches tall.

Now, let q be the proposition that the tree is $i+1$ inches tall. By (1_i), Mr Magoo knows $q \supset \sim(2_i)$; by (3_i), he knows (2_i). Now, $\sim q$ is a logical consequence of $q \supset \sim(2_i)$ and (2_i). Consequently, by (C), (1_i) and (3_i) imply that Mr Magoo knows $\sim q$:

(2_{i+1}) Mr Magoo knows that the tree is not $i+1$ inches tall.

Consequently, from (KK), (C) and (2_i) we can infer (2_{i+1}). By repeating the argument for values of i from 0 to 665, starting from (2_0) we reach the conclusion (2_{666}):

(2_0) Mr Magoo knows that the tree is not 0 inches tall.

(2_{666}) Mr Magoo knows that the tree is not 666 inches tall.

Statement (2_{666}) is false, for the tree is 666 inches tall and knowledge is factive. Thus, given the premises (1_0), . . ., (1_{665}), (2_0), (C), and (KK), we can deduce the false conclusion (2_{666}). Therefore, at least one of (1_0), . . ., (1_{665}), (2_0), (C), and (KK) is to be rejected. Premise (1_i) has already been defended for all i, and (2_0) is obviously true. Consequently, either (C) or (KK) is to be rejected.

Could we reject the assumption (C) that Mr Magoo's knowledge of the pertinent propositions is deductively closed? Assumption (C) is true if deduction is a way of extending one's knowledge: that is, if knowing $p_1, \ldots, p_n$, competently deducing q, and thereby coming to believe q is in general a way of coming to know q. Call that principle *intuitive closure*. Since by hypothesis Mr Magoo satisfies the conditions for the intuitive closure principle to apply, rejecting (C) is tantamount to rejecting intuitive closure. Robert Nozick's counterfactual analysis of knowledge is famously inconsistent with intuitive closure, but that is usually taken as a reason for rejecting the analysis, not for rejecting closure. Chapter 7 will provide arguments against counterfactual conditions on knowledge even of quite a weak kind; a fortiori they are arguments against Nozick's analysis.

A different objection occasionally made to intuitive closure is that even if one's premises are individually probable enough to count as known, one's conclusion might not be. For a logical consequence of several propositions may be less probable than each of them. If there are a million tickets in the lottery and only one wins, each proposition of the form 'Ticket *i* does not win' has a probability of 0.999999, yet the conjunction of all those propositions has a probability of 0. But that objection misconceives the relation between probability and knowledge; however unlikely one's ticket was to win the lottery, one did not know that it would not win, even if it did not (see also section 11.2). No probability short of 1 turns true belief into knowledge. Chapter 10 provides a very different understanding of the connection between knowledge and probability; it does not threaten intuitive closure.

The appeal to probability is in any case unavailing, for the argument can be reworked so that (C) is applied only to single-premise inferences; if q is a logical consequence of p then q is at least as probable as p. For the considerations that supported (1_i) also support:

(4_0) Mr Magoo knows that (for all natural numbers m (if the tree is $m+1$ inches tall then he does not know that it is not m inches tall) and (the tree is not 0 inches tall)).

Parentheses have been inserted to clarify scope. Now suppose, for some given i:

(4_i) Mr Magoo knows that (for all natural numbers m (if the tree is $m+1$ inches tall then he does not know that it is not m inches tall) and (the tree is not i inches tall)).

By (KK) we have:

(5_i) Mr Magoo knows that he knows that (for all natural numbers *m* (if the tree is *m*+1 inches tall then he does not know that it is not *m* inches tall) and (the tree is not *i* inches tall)).

But Mr Magoo knows with certainty that if he knows a conjunction then the first conjunct is true and he knows the second. Thus:

(6_i) Mr Magoo knows that (for all natural numbers *m* (if the tree is *m*+1 inches tall then he does not know that it is not *m* inches tall) and (he knows that the tree is not *i* inches tall)).

But (C) for single-premise deductions applied to (6_i) gives:

(4_{i+1}) Mr Magoo knows that (for all natural numbers *m* (if the tree is *m*+1 inches tall then he does not know that it is not *m* inches tall) and (the tree is not *i*+1 inches tall)).

The inference from (4_i) to (4_{i+1}) is the required sorites step. If we iterate it for each *i* from 0 to 665, starting with (4_0), we reach:

(4_{666}) Mr Magoo knows that (for all natural numbers *m* (if the tree is *m*+1 inches tall then he does not know that it is not *m* inches tall) and (the tree is not 666 inches tall)).

Statement (4_{666}) is false, for the tree is 666 inches tall. Thus the problem does not depend on applying (C) to deductions with more than one premise.

We should in any case be very reluctant to reject intuitive closure, for it *is* intuitive. If we reject it, in what circumstances can we gain knowledge by deduction? Moreover, the closely related anti-luminosity argument in section 4.3 did not assume closure in any form, which suggests that it is not the crucial premise.

A different objection to the argument is that vagueness is somehow to blame. Section 4.5 discussed the same objection. Since the reasons for dismissing it are the same as before, they will not be repeated in detail here. The crucial point is that the premises of the argument are not justified by vagueness in 'know' but by limits on Mr Magoo's eyesight and his knowledge of them. In checking that (1_i) remains true when 'know' is sharpened, we must be careful because 'know' occurs twice in (1_i), which ascribes to Mr Magoo knowledge that he could express in the words 'If the tree is *i*+1 inches tall, then I do not know that the tree is not *i* inches tall'. But if we sharpen 'know' by stipulating a high standard for its application, we make that conditional harder to falsify and therefore easier to know, because the only occurrence of 'know' in the sentence is negative. Since (1_i) was clearly true prior to the sharpening, it therefore remains true afterwards; we may legitimately assume that Mr

Magoo has considered the sharpened sense of 'know'. That will not improve his eyesight. The argument does not rely on the vagueness of 'know'.

Given (C) and (KK) as auxiliary premises, there is a valid argument with otherwise true premises and a false conclusion. Premise (C) is accepted. Therefore, (KK) is to be rejected. Mr Magoo knows something pertinent without knowing that he knows it. Since (KK) follows from the assumption that the condition that one knows a proposition is luminous and background assumptions about Mr Magoo, the luminosity assumption is false. As in section 4.5, we can check that rejecting luminosity really does meet the difficulty by constructing a formal model of (C), $(1_0), \ldots, (1_i), \ldots, (2_0)$ and the negation of (2_{666}) (Appendix 2 has more details).

Mr Magoo cannot identify the particular proposition for which (KK) fails. In general, one cannot knowingly identify a particular counterexample to the KK principle in the first person present tense. If I know that I both know p and do not know that I know p, I must know the first conjunct of that conjunction (since knowing a conjunction entails knowing its conjuncts), that is, I must know that I know p, so the second conjunct is false, so I do not know the conjunction after all (since knowledge is factive); Chapter 12 discusses this kind of argument in more depth. The point may help to explain the seductiveness of the KK principle.

The crucial features of the example are common to virtually all perceptual knowledge. Thus the argument generalizes to show that our knowledge is pervaded by failures of the KK principle. To the informed observer, hearing gives some knowledge about loudness in decibels, and touch about heat in degrees centigrade. When I smell the milk I have some knowledge of the number of minutes since it was opened; when I taste the tea I have some knowledge of how many grains of sugar were put in. The point generalizes to knowledge from sources beyond present perception, such as memory and testimony. This is partly because they pass on inexact knowledge originally derived from past perception, partly because they add further ignorance themselves. How long was my last walk in steps? How long was someone else's walk, described to me as 'quite long'? In each case the possible answers lie on a scale, which can be divided so finely that if a given answer is in fact correct, then one does not know that its neighbouring answers are not correct, and one can know that one's powers of discrimination have that limit. The argument then proceeds as in the case of the distant tree.[1]

[1] The argument of section 5.1 is similar in form to the argument used by Nathan Salmon (1982: 238–40, 1986, 1989) against the S4 principle $\Box p \supset \Box\Box p$ for metaphysical

5.2 FURTHER ITERATIONS

We can generalize the argument of section 5.1 to further iterations of knowledge. We define them inductively. One knows^0 *p* if and only if *p* is true. For any natural number *k*, one knows^{k+1} *p* if and only if one knows^k that one knows *p*. To know^1 *p* is to know *p*, to know^2 *p* is to know that one knows *p*, and so on.

For any *k*, we can argue in parallel with section 5.1 that one can know^k something without being in a position to know that one knows^k it. For if we make suitably modified assumptions about the height and distance of the tree, Mr Magoo's eyesight, his knowledge of its limitations, and his powers of reflection, we can construct a situation in which these modified assumptions are true for a given *k* and all *i*:

(1_i^k) Mr Magoo knows^k that if the tree is $i+1$ inches tall, then he does not know that the tree is not *i* inches tall.

(2_0^k) Mr Magoo knows^k that the tree is not 0 inches tall.

(C^k) If *p* and all members of the set X are pertinent propositions, *p* is a logical consequence of X, and Mr Magoo knows^k each member of X, then he knows^k *p*.

Now make these two assumptions, for a given number *i*:

(2_i^k) Mr Magoo knows^k that the tree is not *i* inches tall.

(KK^k) For any proposition *p*, if Mr Magoo knows^k *p* then he knows^{k+1} *p*.

Since knowing^{k+1} is equivalent to knowing^k that one knows, (2_i^k) and (KK^k) entail:

(3_i^k) Mr Magoo knows^k that he knows that the tree is not *i* inches tall.

Assumptions (1_i^k), (3_i^k), and (C^k) entail:

necessity. The form is valid; the question in each case is whether the premises are plausible. Williamson 1990a: 129 suggests that the plausibility of one of Salmon's premises comes from a source that generates sorites paradoxes; Salmon 1993 disagrees. It is hard to adjudicate disputes about whether intuitions have a common source. The structure of Salmon's premise in itself does not commit him to a sorites paradox. On the other hand, if the validity of the S4 principle is built into our conception of unrestricted metaphysical possibility and necessity, those who find the major premises of sorites paradoxes plausible would also be likely to find Salmon's premise plausible. The analogue of Salmon's argument may be sound for the restricted notions of possibility and necessity discussed in section 5.3, which might also help to account for the plausible appearance of the premises in Salmon's original version.

($2_{i+1}{}^{k}$) Mr Magoo knows^{k} that the tree is not i+1 inches tall.

Suppose that the tree is in fact n inches high. By repeated application of the argument from ($2_{i}{}^{k}$) to ($2_{i+1}{}^{k}$), starting with $2(_{0}{}^{k})$, we reach:

($2_{n}{}^{k}$) Mr Magoo knows^{k} that the tree is not n inches tall.

Since knowledge^{k} is as factive as knowledge, ($2_{n}{}^{k}$) is false. It was deduced from the assumptions ($1_{0}{}^{k}$), . . ., ($1_{n-1}{}^{k}$), ($2_{0}{}^{k}$), (C^{k}), and (KK^{k}). By construction of the example, ($1_{0}{}^{k}$), . . ., ($1_{n-1}{}^{k}$), ($2_{0}{}^{k}$), and (C^{k}) are true; therefore (KK^{k}) is false. The replies to objections to the argument follow the pattern of section 5.1. Thus one can know^{k} something without being in a position to know^{k+1} it. In other words, one can know^{k} something without being in a position to know that one knows^{k} it.

By contrast, some other objections to the general KK thesis do not threaten the corresponding generalization of (KK^{k}) for k>1. For example, a simple creature might know that it was snowing without knowing that it knows that it was snowing because the latter, unlike the former, requires it to have a concept of knowledge, which it lacks. But if k≥2 and one knows^{k} p, then one knows something concerning knowledge and so has the concepts needed for knowing^{k+1} p.

Can we combine all finite iterations of knowledge? One knows$^{\omega}$ p if and only if for every natural number k one knows^{k} p. Can we mimic the foregoing argument with ω in place of k? The premises of the reductio ad absurdum are these:

($1_{i}{}^{\omega}$) Mr Magoo knows$^{\omega}$ that if the tree is i+1 inches tall, then he does not know that the tree is not i inches tall.

($2_{0}{}^{\omega}$) Mr Magoo knows$^{\omega}$ that the tree is not 0 inches tall.

(C^{ω}) If p and all members of the set X are pertinent propositions, p is a logical consequence of X, and Mr Magoo knows$^{\omega}$ each member of X, then he knows$^{\omega}$ p.

(KK^{ω}) For any proposition p, if Mr Magoo knows$^{\omega}$ p then he knows$^{\omega}$ that he knows p.

For some n, the false conclusion is this:

($2_{n}{}^{\omega}$) Mr Magoo knows$^{\omega}$ that the tree is not n inches tall.

We might conclude on the basis of ($1_{0}{}^{\omega}$), . . ., ($1_{n-1}{}^{\omega}$), ($2_{0}{}^{\omega}$), and (C^{ω}) that Mr Magoo is a counterexample to (KK^{ω}). But that is the wrong moral to draw from this example, for (KK^{ω}) is a logical truth. If Mr Magoo knows$^{\omega}$ p, then for each natural number k he knows^{k+1} p, which is to know^{k} that he knows p, so he knows$^{\omega}$ that he knows p. Thus ($1_{0}{}^{\omega}$), . . .,

(1_{n-1}^{ω}), (2_0^{ω}), and (C^{ω}) entail the false conclusion (2_n^{ω}) by themselves; one of them is false. Given a natural number k, we can construct an example in which (1_0^k), . . ., (1_{n-1}^k), and (2_0^k) are true, by finite adjustments of the original case, which are clearly possible. An infinite adjustment turns out to be impossible. That does not undermine the morals drawn from the earlier versions of the argument. The crude point is that iterating knowledge is hard, and each iteration adds a layer of difficulty. Knowledge$^{\omega}$ involves infinitely many layers of difficulty. Under some conditions, that amounts to impossibility. The next section develops these remarks more systematically.

Knowledge$^{\omega}$ presents an interesting challenge to the generalized argument against luminosity in Chapter 3. Since it seems possible in principle to gain or lose knowledge$^{\omega}$, one might expect the argument to show that one can know$^{\omega}$ without being in a position to know that one knows$^{\omega}$. But that conclusion is problematic. For if one knows$^{\omega}$ *p*, then one knows each member of the set containing the proposition that one knows^k *p* for each natural number *k*; thus one knows the premises of a deductively valid argument to the conclusion that one knows$^{\omega}$ *p*; one is therefore in some sense in a position to know that one knows$^{\omega}$ *p*. The condition that one knows$^{\omega}$ *p* seems to be luminous.

The argument might be challenged on the grounds that we are not in a position to make inferences with infinitely many premises. Indeed, even when an inference has only finitely many premises, it is not obvious that we are always in a position to know that which follows deductively from what we know. Only in a rather attenuated sense are we in a position to know all the consequences of the axioms of Peano Arithmetic. However, this response is not wholly satisfying, for the original argument against luminosity made no appeal to limits on powers of inference. If the condition that one knows$^{\omega}$ *p* is luminous in the attenuated sense, why does the original argument not generalize to this case?

Knowing$^{\omega}$ may fail the gradualness requirement. Although someone can gain or lose knowledge$^{\omega}$, the change may necessarily be sudden. After all, it is the change from finitely many iterations of knowledge to infinitely many or vice versa; how could it be gradual? If knowing$^{\omega}$ does fail the gradualness requirement, it will be a hard state to enter or leave: how is one to jump instantaneously from the finite to the infinite or back again? The kind of common knowledge that we are supposed to have of conventions is usually defined in a way that requires us to know$^{\omega}$. For example, if John knows that Jane knows that John knows that Jane knows that John knows *p*, then John knows that John knows that John knows *p*, if he is sufficiently reflective. Common knowledge would therefore be a convenient idealization, like a frictionless plane.

The convenience need not be confined to the theoretician. Perhaps some everyday practices of communication and decision-making depend on a pretence that we have common knowledge. That hardly comes as a surprise, for infinitely many of the propositions involved in common knowledge are too complex for humans to be psychologically capable of entertaining them. The present point is that the obstacles to entertaining them are not the only obstacles to knowing them.[2]

5.3 CLOSE POSSIBILITIES

A reliability condition on knowledge was implicit in the argument of section 5.1 and explicit in sections 4.3 and 4.4. We have seen that such a condition generates an obstacle to iterating knowledge. We can better understand the nature of the obstacle by considering reliability in the more general context of a family of related notions such as safety, stability, and robustness.

Imagine a ball at the bottom of a hole, and another balanced on the tip of a cone. Both are in equilibrium, but the equilibrium is stable in the former case, unstable in the latter. A slight breath of wind would blow the second ball off; the first ball is harder to shift. The second ball is in danger of falling; the first ball is safe. Although neither ball did in fact fall, the second could easily have fallen; the first could not. The stable equilibrium is robust; the unstable equilibrium, fragile.

Reliability and unreliability, stability and instability, safety and danger, robustness and fragility are modal states. They concern what could easily have happened. They depend on what happens under small variations in the initial conditions. If determinism holds, it follows from the initial conditions and the laws of nature that neither ball falls. But it does not follow that both balls were in stable equilibrium, safe from falling, for the initial conditions themselves could easily have been slightly different. There is a danger in a given case that an event of type E will occur (for example, that the ball will fall) if and only if in some sufficiently similar case an event of type E does occur. The danger is slight if E occurs in very few sufficiently similar cases, but that is not the same as a distant danger, which occurs only in insufficiently similar

[2] Fagin, Halpern, Moses, and Vardi 1995: 395–422 discuss some weakenings of the notion of common knowledge. In general, forms of almost-common knowledge do not imply almost the same behaviour as common knowledge itself, which raises difficult problems beyond the scope of this book. Shin and Williamson 1996 discuss common belief in a context of inexact knowledge.

cases. The relevant similarity is in the initial conditions, not in the final outcome (with the laws presumably held fixed). 'Initial' here refers to the time of the case, not to the beginning of the universe; I may be safe once I have caught the last flight out of the besieged city, even though I could easily have been a few minutes late and missed the flight, in which case I should now have been in danger. Safety and danger are highly contingent and temporary matters. Just how similar the case must be to one in which an event of type E occurs for the term 'danger' to apply depends on the context in which the term is being used.[3]

Reliability resembles safety, stability, and robustness. These terms can all be understood in several ways, of course. For present purposes, we are interested in a notion of reliability on which, in given circumstances, something happens reliably if and only if it is not in danger of not happening. That is, it happens reliably in a case α if and only if it happens (reliably or not) in every case similar enough to α. In particular, one avoids false belief reliably in α if and only if one avoids false belief in every case similar enough to α. When the danger is a matter of degree, reliability involves a trade-off between the degree to which the danger is realized and the closeness of the case in which it is realized. A very high degree of realization in a not very close case and a lower degree of realization in a closer case both make for unreliability. The argument of section 4.3 involved such a trade-off, the closeness of case α_{i+1} to case α_i compensating for the slightly lower degree of belief in α_{i+1}.

On a topological conception, a point x counts as safely in a region R if and only if x is in the interior of R. If R is a region in a metric space defined by some real-valued measure of distance, x is in the interior of R if and only if for at least one positive real number c, every point whose distance from x is less than c belongs to R. More generally, x belongs to the interior of R if and only if x belongs to some open subset of R. There is no difficulty in iterating safety on this conception, for the interior of the interior of R is just the interior of R. Thus x is safely safely in R—that is, safely in the region that contains all and only the points that are safely in R—if and only if x is safely in R. For if x is safely in R, then, for some non-zero distance c, every point less than c from x is in R, so every point less than $c/2$ from a point less than $c/2$ from x is in R, so every point less than $c/2$ from x is safely in R, so x is safely safely in R. On a corresponding conception of stability, a ball balanced in an indentation on the tip of the cone is in stable equilibrium, no matter how small and shallow the indentation.

For most practical purposes, the topological conception is not the

[3] See Sainsbury 1997 and Peacocke 1999: 310–28 for more discussion of the notion of easy possibility. It is applied in Williamson 1994b: 226–30.

one we need. The indentation must be of a certain size and depth for the ball not to be blown off by prevalent light breezes. To be safe on the top of a cliff, a young child must be at least three feet from the edge; it is not enough to be some positive distance or other, no matter how small, from the edge. Naturally, features of the context may contribute to fixing the margin for something to count as 'safe': for example, the severity of the consequences if one succumbs. Suppose that in some context a point is safely in a region if and only if every point less than three feet away is in the region. Then a point can be safely in a region R without being safely safely in R, for if the nearest point to x not in R is four feet away, x is safely in R but only two feet from a point two feet from a point not in R, so x is two feet from a point not safely in R, so x is not safely safely in R. The notion of what could easily happen behaves like the dual of safety; 'It could easily have been F' is close to 'It was not safely not F'. If it could easily have happened that an event of type E could easily have happened, it does not follow that an event of type E could easily have happened. For example, if exactly i humans were now alive, then it would be the case that it could easily have happened that exactly i+1 humans were now alive, but for some sufficiently large number k it would not be the case that it could easily have happened that exactly i+k humans were now alive. If the actual number is i, then it could easily have happened that it could easily have happened . . . [k times] . . . that exactly i+k humans were alive now, but it could not easily have happened that exactly i+k humans were alive now. Thus iterations of 'it could easily have happened that' do not collapse.

The failures of knowledge to iterate observed in sections 5.1 and 5.2 are closely related to the failure of safety and reliability to iterate. One can be safe without being safely safe. In particular, one can be safe from error without being safely safe from error. One can be reliable without being reliably reliable. Since knowledge requires reliability, it is hardly surprising that one can know without knowing that one knows.

Safety is hard to iterate. For each natural number k, we can define x to be safely^k in R if and only if x is safely safely . . . [k times] . . . in R; x is safely$^\omega$ in R if and only if x is safely^k in R for every natural number k. Suppose that for some fixed non-zero distance c, a point is safely in a region if and only if every point less than c from the point is in the region. In n-dimensional Euclidean space, any two points are linked by a finite sequence of intermediate points each less than c from the next. Thus, unless R is the whole space, no point is safely$^\omega$ in R. A luminous condition resembles a region every point in which is safely in it; consequently, every point in such a region is safely$^\omega$ in it. In this instance, the only such regions of Euclidean space are the whole space and the null region. Similarly, we might think of a formula A as luminous in a system of epistemic logic if and only if $A \supset KA$ is a theorem. The analogous feature would then be that $A \supset KA$ is a theorem only if either A is a theorem (A corresponds to a region that is the whole space) or its negation is a theorem (A corresponds to the null region). Some natural systems have that property (see Appendix 2 and Williamson 1992a).

If R is the complement in full Euclidean space of a non-null bounded region (a sphere, for example), then for every natural number k some points are safely^k in R, even though no point is safely$^\omega$ in R. But if R itself is a bounded region, then for some natural number k no point is even safely^k in R.

Euclidean space is not the only kind of space, of course. We should not assume without argument that the space of possibilities in which we are interested has a Euclidean structure. In principle, it might consist of several disconnected regions. Every point in one of those regions might be safely in it; consequently, every point in the region is safely$^\omega$ in it. We also cannot assume that the required margin for safety c is uniform throughout the space. Prevailing winds may be stronger in some areas than in others. If they have a prevailing direction, one may be more easily blown from x to y than from y to x. Suppose, for example, that the closer one comes to a fixed point z_0 the more conditions favour stability. We can imagine contexts in which the required margin for safety at each point is its distance from z_0. Thus, unless x is z_0 itself, any point y is easily accessible from x if and only if y is closer to x than z_0 is; x is safely in a region R if and only if every point easily accessible from x is in R. If we fix a margin for safety at z_0 too, every point has a margin for safety. But since z_0 is accessible from no point other than itself, every point in the region consisting of the whole space except for z_0 is safely in that region. Thus every point in that region is safely$^\omega$ in it. Formally, such examples model non-trivial luminous conditions. Chapter 4 indicates that such a model would not be an accurate representation of knowledge.

Suppose that one is in a position to know only if one is safe from error in the relevant respect. We might try to deduce that, if a condition C can obtain without safely obtaining, then C can obtain even if one is not in a position to know that C obtains, and therefore that C is not luminous. The idea would be that one is in a position to know that C obtains only if one is safe from error in believing that C obtains, which requires C to obtain safely. But that is too quick. To be safe from error in believing that C obtains is to be safe from falsely believing that C obtains. Thus in a case α one is safe from error in believing that C obtains if and only if there is no case close to α in which one falsely believes that C obtains. But even if in α one believes that C obtains and is safe from error in doing so, it does not follow that C obtains in every case close to α, for there may be cases close to α in which C does not obtain and one does not believe that it obtains. One can believe that C obtains and be safe from error in doing so even if C does not safely obtain, if whether one believes is sufficiently sensitive to whether C obtains. For example, one may be safe from error in believing that the child is not falling even though she is not safe from falling, if one is in a good position to see her but not to help her.

We need a further assumption to generate an argument against luminosity. If we combine the safety from error requirement on knowledge with limited discrimination in the belief-forming process and some plausible background assumptions, then we can deduce failures of luminosity. That is not intended to formalize the anti-luminosity argument of Chapter 4, which depends on applying reliability considerations in a subtler way to degrees of confidence. The argument below models those considerations under highly simplified assumptions, which permit us to restrict our attention to the binary contrast between believing and not believing. It explains how the model falsifies luminosity and verifies a margin for error principle.

Suppose that for some parameter v, such as the height of the tree, for every case α, whether the condition C obtains in α depends only on the value $v(\alpha)$ of v in α. For example, C might be the condition that the tree is at most fifty feet high. We may assume for simplicity that v takes non-negative real numbers as values. To be explicit:

(7) For all cases α and β, if $v(\alpha) = v(\beta)$ then C obtains in α if and only if C obtains in β.

In many examples, something like the following will hold, for some small positive real number c:

(8) For all cases α and non-negative real numbers u, if $|u - v(\alpha)| < c$ and in α one believes that C obtains then, for some case β close to α, $v(\beta) = u$ and in β one believes that C obtains.

Less formally: if one has the belief, then one could easily still have had it if the parameter had taken a given slightly different value. One's belief is not perfectly discriminating. As already noted, iterations of close possibility do not collapse, so (8) does not entail that, if one has the belief, then one could easily still have had it if the parameter had taken a very different value. If one believes that the tree is at most fifty feet high, then one could easily still have believed that if the tree had been an inch higher, but not if it had been one hundred feet higher.

Now assume a connection between knowledge and safety from error:

(9) For all cases α and β, if β is close to α and in α one knows that C obtains, then in β one does not falsely believe that C obtains.

In a more careful version of (9), we might qualify both 'know' and 'believe' by 'on a basis B'. Knowledge on one basis (for example, seeing an event) is quite consistent with false belief in a close case on a very different basis (for example, hearing about the event). We might also relativize (8) and (9) to a subclass of cases by restricting the quantifiers over cases to that subclass. The argument below will still go through if we modify the other propositions in the same way. For simplicity, we may ignore these complications.

We must also articulate a connection between knowing and being in a position to know. One is in a position to know something determined by the value of a parameter only if one can know without changing the value of the parameter:

(10) For all cases α, if in α one is in a position to know that C obtains then, for some case β, $v(\alpha) = v(\beta)$ and in β one knows that C obtains.

Statement (10) can be understood as a stipulation about the meaning of 'in a position to know'.

Finally, we assume that knowledge implies belief:

(11) For all cases α, if in α one knows that C obtains then in α one believes that C obtains.

From (7)–(11) and the assumption (L) that C is a luminous condition, we can deduce this:

(12) For all cases α and β, if $|v(\alpha)-v(\beta)|<c$ then C obtains in α if and only if C obtains in β.

For suppose that C obtains in α and $|v(\alpha)-v(\beta)|<c$. By (L), in α one is in a position to know that C obtains. By (10), for some case α^*, $v(\alpha) = v(\alpha^*)$ and in α^* one knows that C obtains. Thus $|v(\alpha^*)-v(\beta)|<c$ and, by (11), in α^* one believes that C obtains. Consequently, by (8), for some case β^* close to α^*, $v(\beta^*) = v(\beta)$ and in β^* one believes that C obtains. Since β^* is close to α^* and in α^* one knows that C obtains, by (9) in β^* one does not falsely believe that C obtains. Therefore, C obtains in β^*. Since $v(\beta^*) = v(\beta)$, C obtains in β by (7). This shows that if $|v(\alpha)-v(\beta)|<c$ then C obtains in α only if C obtains in β. The converse is similar.

Statement (12) is a disastrous conclusion if the parameter v can vary continuously in this sense:

(13) For all non-negative real numbers u, for some case α, $v(\alpha) = u$.

For (12) and (13) entail:

(14) For all cases α and β, C obtains in α if and only if C obtains in β.

For any real number can be reached from any other in a series of arbitrarily short steps; there will be a sequence of non-negative real numbers $u_0, \ldots, u_n$ such that $u_0 = v(\alpha)$, $u_n = v(\beta)$, and, for all i, $(0 \leq i < n)$, $|u_i - u_{i+1}| < c$. By (13), there is a corresponding sequence of cases $\alpha_0, \ldots, \alpha_n$ such that $v(\alpha_i) = u_i$ for all i $(0 \leq i \leq n)$, where $\alpha_0 = \alpha$ and $\alpha_n = \beta$. Consequently, for all i $(0 \leq i < n)$, $|v(\alpha_i) - v(\alpha_{i+1})| < c$, so, by (12), C obtains in α_i if and only if C obtains in α_{i+1}. By the transitivity of the biconditional, C obtains in α if and only if C obtains in β. Thus C obtains in all cases or in none; it is trivial. Contrapositively, if C is not trivial and the assumptions (7)–(11) and (13) hold, then C is not luminous.

If we like, we can replace the assumption (13) that the parameter v varies continuously by the weaker assumption that v varies in an approximately continuous way, in the sense that for every non-negative real number u there is a case α such that $|u - v(\alpha)| < c/3$.

When we drop the luminosity assumption (L), we can still deduce this consequence from (7)–(11):

(15) For all cases α and β, if $|v(\alpha) - v(\beta)| < c$ and in α one is in a position to know that C obtains then C obtains in β.

The argument for (15) is like the argument for (12), but without the initial application of (L). Statement (15) is a *margin for error* principle: one knows that a condition obtains only if it obtains in all cases in which the relevant parameter differs at most slightly in value. The disastrous conclusion (14) that C is trivial follows easily from (13), (15), and (L). Since (13) or a suitable weakening of it is usually uncontentious, the margin for error principle usually blocks luminosity.

The margin for error c may depend on the condition C. However, if conditions $C_0, \ldots, C_k$ satisfy (15) with respect to margins for error $c_0, \ldots, c_k$ respectively (for the same parameter v), then of course $C_0, \ldots, C_k$ all satisfy (15) with respect to the minimum of $c_0, \ldots, c_k$. But an infinite class of conditions each with a positive margin for error might not have a common positive margin for error, for the greatest lower bound of their individual margins for error might be 0.

The argument of this section does not justify us in believing that every condition satisfies a principle like (15). The argument for (15) depends on the premise (8), that one's belief is not perfectly discriminating with respect to the underlying parameter. That assumption is not obvious, especially if the underlying parameter itself constitutively depends on one's belief, as some philosophers postulate for phenomena that they would classify as response-dependent. For example, they hold that the intensity of one's pain constitutively depends on one's beliefs about the intensity of one's pain. Such cases require the subtler argument of Chapter 4. Nevertheless, the assumptions (7)–(11) and (13) are plausible in a wide range of cases; they explain margins for error and the failure of luminosity. In particular, if the condition that one knows (or that one knows^k) that C obtains satisfies anything like (15) in place of C—naturally, with a parameter v that encodes enough about the case to determine whether one knows—then one will expect just the kind of difficulty in iterating knowledge that sections 5.1 and 5.2 observed. In particular, the crucial premises (1_i) and (1_i^k) simply attribute to Mr Magoo knowledge of a contraposed instance of the margin for error principle (15). Every iteration requires a further margin.

5.4 POINT ESTIMATES

I might reach my belief about the height of a tree by estimating its height and then applying an upper bound on the inaccuracy of my estimate.[4] For example, I estimate that the tree is 55 feet high, and come to believe that it is between 50 and 60 feet high, on the grounds that in these circumstances my estimate will not be out by more than 5 feet. In effect, I deduce (18) from the premises (16) and (17):

(16) I estimated that the tree is fifty-five feet high.

(17) My estimate of the height of the tree differs from the height of the tree by at most five feet.

(18) The tree is between fifty and sixty feet high.

Since I reached the conclusion (18) by inference from (16) and (17), I know (18) if and only if I know (16) and (17). Suppose that in these circumstances my estimates are never out by more than five feet, but are sometimes out by as much as five feet. Thus I might estimate that the

[4] Sections 5.4 and 5.5 answer points raised by Peter Mott 1998 and others. Williamson 2000b considers some further details of Mott's arguments.

tree is fifty-five feet high when it is in fact fifty feet high. In that case, it may appear, I can know (18) without satisfying any principle like (15). If the tree were even slightly less tall, my belief would be false.

The objection assumes that I can know that my estimate was out by at most five feet when it was in fact out by exactly five feet. That is in effect to assume that I need no margin for error in my knowledge of the accuracy of my own estimates. But my belief about my own accuracy has no more exact basis than my perceptual beliefs. If I were further away from the tree, or the light were worse, my estimate could be out by more than five feet. My judgement that my estimate of the height of this tree is out by at most five feet depends on my perceptual beliefs about my distance from the tree and the quality of the light. If I believe that my estimate is out by at most five feet, when in fact it is out by exactly five feet, then I could easily have formed that belief in slightly different circumstances in which my estimate was out by slightly more than five feet. Certainly the objector has not shown that one can know in the envisaged circumstances that one's estimate is out by at most five feet when in fact it is out by exactly five feet. There is almost no limit to how far out my estimates can be on a really bad day.

If my estimate is more than five feet out, I cannot know that it is at most five feet out, simply because knowledge is factive. A margin for error principle exhibits a further way in which my knowledge of the accuracy of my estimate depends on the accuracy of that estimate.

Naturally, we can imagine situations in which one knows exactly how far out one's estimate can be, just as we can imagine situations in which one knows exactly how tall the tree is. But those situations involve ways of knowing quite different from those we actually employ. The objector has done nothing to show that our actual methods enable us to dispense with margins for error. When C is the condition that one's estimate is out by at most five feet, the premises of the argument for (15) remain plausible. If one's knowledge of upper bounds on the inaccuracy of one's estimate of the height of the tree satisfies margin for error principles, then one's derivative knowledge of the height of the tree will satisfy a corresponding margin for error principle.

5.5 ITERATED INTERPERSONAL KNOWLEDGE

Iterating knowledge is hard, whether it is knowing about one's own knowledge or knowing about another's. Do margin for error principles make it too hard?

Imagine Steven, Anna, and John looking at a tree. It is in fact fifteen feet high. They satisfy a margin for error principle with a margin of five feet:

(19) If Steven, Anna or John knows that the tree is at most n+5 feet high then the tree is at most n feet high.

Statement (19) depends on their eyesight and visual judgement, the distance of the tree and the quality of the light. Steven judges out loud that the tree is at most twenty-five feet high. Anna hears him and judges out loud that Steven knows that the tree is at most twenty-five feet high. John hears Anna and Steven and judges that Anna knows that Steven knows that the tree is at most twenty-five feet high.

We might be tempted to argue that (19) implies the implausible restriction on common knowledge that John does not know that Anna knows that Steven knows that the tree is at most twenty-five feet high. For, by (19), if Steven knows that the tree is at most twenty-five feet high then it is at most twenty feet high. Given that Anna knows that conditional and that her knowledge is closed under deduction, if Anna knows that Steven knows that the tree is at most twenty-five feet high then Anna knows that it is at most twenty feet high. By (19), if Anna knows that the tree is at most twenty feet high then the tree is at most fifteen feet high. Consequently, if Anna knows that Steven knows that the tree is at most twenty-five feet high then it is at most fifteen feet high. Given that John knows that conditional and that his knowledge is closed under deduction, if John knows that Anna knows that Steven knows that the tree is at most twenty-five feet high then John knows that it is at most fifteen feet high. By (19), if John knows that the tree is at most fifteen feet high then it is at most ten feet high. Consequently, if John knows that Anna knows that Steven knows that the tree is at most twenty-five feet high, then it is at most ten feet high. But by hypothesis the tree is fifteen feet high, so John does not know that Anna knows that Steven knows that the tree is at most twenty-five feet high.

The argument applies (19) to Anna's knowledge when she has heard Steven and to John's knowledge when he has heard Anna. Thus we must read (19) as describing the knowledge that Steven, Anna, and John have once they have considered the others' judgements, not their unaided knowledge. That consideration might reduce the required margin for error.

The argument also tacitly assumes that John, Anna, and Steven have common knowledge of their final margins for error. If Anna does not know (19), because for all she knows Steven is much better at judging heights than he really is, then Anna may know that Steven knows that the tree is at most twenty-five feet high without herself knowing that it is at most twenty feet high. Similarly, even if both Anna and John do know (19), John may not know that Anna knows (19); perhaps John does not know how well Anna is acquainted with Steven. So although we can argue, given deductive closure and Anna's knowledge of (19), to the conclusion that if Anna knows that Steven knows that the tree is at most twenty-five feet high then it is at most fifteen feet high, in those circumstances John may know (19) without knowing that conditional. Thus even if John knows that Anna knows that Steven knows that the tree is at most twenty-five feet high, John may not know that it is at most fifteen feet high, contrary to the objector's argument.

The objector assumes something like common knowledge of (19). Since margin for error principles undermine much purported common knowledge, an argument against them cannot legitimately assume common knowledge of (19). Steven, Anna and John do not merely have limited knowledge of the height of the tree; they have limited knowledge of the limits on their own and others' knowledge of the height of the tree. For they have limited knowledge of their eyesight and visual judgement, the distance of the tree, and the quality of the light. Their second-order knowledge may therefore be expected to satisfy further margin for error principles such as this:

(20) If Anna knows that Steven's margin for error for the height of the tree is at least i feet then Steven's margin for error for the height of the tree is at least $i+j$ feet.

But (19) does not entail that (20) holds when $j = 5$; the required value may be smaller. To make (20) rigorous, we should define just what it means to speak of someone's margin for error; that can be done in more than one way, but something like (20) will hold on all of them.

Steven, Anna, and John may spend so much time together that they know each other as judges of height as well as they know themselves. That reduces the intersubjective case to the intrasubjective case. Analogous considerations apply. Even if I conduct experiments to test my reliability, my knowledge of my own margins for error will remain inexact and subject to further margin for error principles. Indeed, the margin for error varies with the height of the tree—it is smaller for trees less than ten feet high—which is just what I am trying to judge. Moreover, since the width of the margin required to satisfy a margin for error principle depends on the vague concept of knowledge, we have no means of measuring margins for error accurately even given the height of the tree. Statement (19) does not entail that Anna cannot know that she knows that she knows that the tree is at most twenty-five feet high.

In the simplest models of margin for error principles, we can treat those principles as common knowledge. We can then derive tight constraints on the number of iterations of knowledge. Such models exhibit some of the main structural features of knowledge subject to margin for error principles. For every natural number *n*, *n* iterations of knowledge are insufficient for *n*+1 iterations. But such models are not intended to be realistic; they embody many oversimplifications. To complain that they have unrealistic consequences is pointless. If we want to give a more realistic model of margins for error, we can do so in various ways discussed in section 5.3 (see also Appendix 2). If the width of the margin varies from point to point, non-trivial conditions can be commonly known to obtain. Nevertheless, the crude moral remains: every iteration of knowledge, intrasubjective or intersubjective, adds a new layer of difficulty. These difficulties manifest themselves in some notorious paradoxes, such as the Surprise Examination, discussed in the next chapter. Section 10.5 generalizes margin for error principles from knowledge to non-factive cognitive notions, such as high probability on one's evidence.

6
An Application

6.1 SURPRISE EXAMINATIONS

We can present a structural analogue of the argument in section 5.1 about the distant tree as a paradox.[1] The Glimpse, as we may call it (#1 below), stands at one corner of a two-dimensional array of paradoxes, with the notorious Surprise Examination (###4) at the diagonally opposite corner. This connection will enable us to appreciate the relevance of the ideas developed in Chapter 5 to the Surprise Examination, and to suggest a solution. The term 'paradox' is not intended to imply insolubility.

Let *n* be the number of days in a school term.

#1. A teacher's pupils know that she rings all and only examination dates on the calendar in her office. At the beginning of term, the only knowledge they have of examination dates this term comes from a distant glimpse of the calendar, enough to see that one and only one date is ringed and that it is not very near the end of term, but not enough to narrow it down much more than that. The pupils recognize their situation. They know now that for all numbers *i*, if the examination is *i*+1 days from the end of term then they do not know now that it will not be *i* days from the end ($0 \leq i < n$). In particular, they know now that if it is on the penultimate day then they do not know now that it will not be on the last day. But they also know now from their glimpse of the calendar that it will not be on the last day. They deduce that it will not be on the penultimate day. They also know now that if it is on the antepenultimate day then they do not know now that it will not be on the penultimate day. They deduce that it will not be on the antepenultimate day. And so on. They rule out every day of term as a possible date for the examination.

[1] Sorensen 1988 has a very full survey of the earlier literature on the Surprise Examination. More recent work includes Bovens 1997, Hall 1999, Jackson 1987, Janaway 1989, Koons 1992: 13–23, Sainsbury 1995: 91–105, and Weintraub 1995.

#2. Like #1, but it is the school caretaker, not the pupils, who catches a glimpse of the calendar. He tells them that the (one and only) ringed date is not very near the end of term. They know him to be a trustworthy observer and informant. In circumstances like these, reliable testimony is a channel for the communication of knowledge. Since the caretaker knows that the ringed date is not very near the end of term, the pupils know it too. They reason as in #1.

#3. Like #2, but it is the teacher, not the caretaker, who gives the pupils information. She tells them just that the examination will not be very near the end of term. They know her to be a trustworthy informant. They reason as in #1.

#4. Like #3, but what the teacher tells the pupils is that, for all *i*, if the examination is *i*+1 days from the end of term then they do not know now that it will not be *i* days from the end ($0 \leq i < n$). What the teacher says is true, for the date she has fixed is not very near the end of term, she is the pupils' only source of information about the date, and what she has told them is only what they worked out for themselves in #1–#3: for all *i*, if the examination is *i*+1 days from the end then they do not know now that it will not be *i* days from the end, for they did not know that in #3, and they know no more about the date in #4 than they did in #3. The teacher knows all this, so she knows the truth of what she tells the pupils. As in #3, they know the truth of what she tells them. They reason as in #1.

##1. Like #1, but the pupils reason slightly differently. They know now that, for all *i*, if the examination is *i*+1 days from the end of term, they will not know then, on morning of the examination, that it will not be *i* days from the end, for they know: it will not be very near the end of term; on no day not very near the end of term will their glimpse of the calendar and their memory of examinationless days enable them to know that it will not be the day after; they will not acquire further relevant information between now and then (by chance there has been a general tightening of school security). In particular, they know that if it is on the penultimate day, then they will not know then that it will not be on the last day. But they will know that then, for they will remember their glimpse of the calendar. They deduce that it will not be on the penultimate day. They also know that if it is on the antepenultimate day, they will not know then that it will not be on the penultimate day. But they will remember the previous conclusion. They deduce that it will not be on the antepenultimate day. And so on. Their conclusion is as in #1.

##2. The pupils' source is as in #2; they reason as in ##1.

##3. The pupils' source is as in #3; they reason as in ##1.

##4. The teacher tells the pupils what in ##1–##3 they worked out for themselves, that, for all *i*, if the examination is *i*+1 days from the end, they will not know then that it will not be *i* days from the end, just as in #4 she tells them what in #1–#3 they worked out for themselves; they reason as in ##1.

###1. Like ##1, but the knowledge the pupils use is that they will not know on the morning of the examination that it will be on that day. For, as in ##1, they will not know that it will not be the day after. They reason as in ##1, except that they need the additional premise that on the morning of the examination they will remember that it was not on any earlier day.

###2. The pupils' source is as in #2; they reason as in ###1.

###3. The pupils' source is as in #3; they reason as in ###1.

###4. The teacher tells the pupils what in ###1–###3 they worked out for themselves: that they will not know on the morning of the examination that it will be on that day. They reason as in ###1.

The twelve arguments differ in two dimensions. The source of knowledge in #1–###1 is perception; in #3–###3 and #4–###4 it is testimony; #2–###2 are mixed cases. The reasoning in #1–#4 concerns present knowledge; in ##1–##4 and ###1–###4 it concerns future knowledge. Nevertheless, the pupils' reasoning is unsound in every case, and the cases are similar enough to make this unlikely to be mere coincidence. A common error should be sought. More specifically, the pupils' reasoning does not seem to get any *better* as one moves from #1 to ###4. If they commit a fallacy in #1, they commit an analogous fallacy in ###4. Arguments ##1–##4 and ###1–###4 are the more complex cases, for the pupils are reasoning about future knowledge on the basis of present knowledge. The gap between what they know at the beginning of term and what they know later leaves room for mistakes that are not at issue in #1–#4, where the pupils reason about present knowledge on the basis of present knowledge. However, their reasoning is unsound in the simpler cases too: their error there is unlikely to have been corrected in the more complex ones. Thus any diagnosis of one or more of ##1–##4 and ###1–###4 which does not extend to #1–#4, although perhaps correct as far as it goes, should be presumed incomplete, not having identified the common error. Since ###4 is the usual paradox of the Surprise Examination, any adequate diagnosis of the Surprise Examination should extend to the Glimpse (#1).

Some diagnoses of the Surprise Examination depend on the gap between what the pupils know at the beginning of term and what they know later, or at least on the epistemic possibility of such a gap (see Wright and Sudbury 1977 and Jackson 1987). Now some variant paradoxes with the same structure concern what different people know at the same time, rather than what the same people know at different times, so the diagnosis is best generalized as depending on any gap between cognitive standpoints (Sorensen 1988: 317–18). The underlying point is the same. In ruling out a last-day examination, the pupils assume that they will still know on the last morning that there will be a surprise examination, defined as an examination on a day when the pupils do not know in the morning that there will be an examination that day. But even if they know that at the beginning of term, they could lose the knowledge later. The point is not that memory is fallible; the pupils may be assumed to know that they will not forget the teacher's announcement. Rather, their memory of examinationless days would undermine their earlier knowledge of the truth of the announcement, like misleading evidence. For them, to know on the last day that there will be a surprise examination, when there has been none so far, is in effect to know 'There will be an examination tomorrow and we do not know that there will be an examination tomorrow'. Such knowledge is impossible, for their knowledge of the first conjunct is inconsistent with the truth of the second (Chapter 12 discusses this kind of reasoning). Thus if the examination is on the last day, then the pupils will have lost their knowledge of the truth of the teacher's announcement by the last morning. Moreover, they can know that conditional in advance. Thus the reasoning by which they rule out a last-day examination is unsound, for it assumes that knowledge will be retained in trying to refute a supposition on which it would not be retained.

The foregoing diagnosis can be elaborated in a variety of ways. There is clearly something to it. Nevertheless, it is incomplete. It yields no objection to the reasoning in the Glimpse, which is an equally unsound simplification of the reasoning in the Surprise Examination. What is wrong in the Glimpse is wrong in the Surprise Examination too, yet unmentioned in the diagnosis.

The analogy with the Glimpse reinforces other points about the Surprise Examination. The teacher's announcement corresponds to the claim in the Glimpse that if the examination is i+1 days from the end, then the pupils do not know that it is not i days from the end. Thus to say that the announcement is false or truth-valueless corresponds to saying that that attempted expression of the content of the pupils' knowledge in the Glimpse is false or truth-valueless. To say that in the Surprise Examination the pupils cannot know in advance that the announcement is true corresponds to saying that in the Glimpse the pupils cannot know the limitation on their knowledge by reflecting on the poverty of their perceptual knowledge in that case. The obvious implausibility of these claims for the Glimpse points up their implausibility for the Surprise Examination too. Advance knowledge that there will be a test, fire drill, or the like of which one will not know the time in advance is an everyday fact of social life, but one denied by a surprising proportion of early work on the Surprise Examination. Who has not waited for the telephone to ring, knowing that it will do so within a week and that one will not know a second before it rings that it will ring a second later? Any adequate diagnosis of the Surprise Examination should allow the pupils to know that there will be a surprise examination.

Other points about the Glimpse generalize to the Surprise Examination. For example, the problem does not depend on self-reference or ungroundedness in the teacher's announcement or the pupils' knowledge of it. For the problem in the Glimpse is not of that kind. Of course, a viciously self-referential twist *can* be given to the teacher's announcement; that does not mean that it *must* be. As noted above, claims like the teacher's are often true, and known to be so. The Glimpse is not a Liar paradox, nor is the Surprise Examination on its natural reading.[2] That vagueness is not to blame is even clearer in the Surprise Examination than in the Glimpse (see also sections 4.5 and 5.1).[3]

The paradox of the Glimpse depends on a concealed use of the KK principle. The pupils know now, at the beginning of term, that if the examination is on the penultimate day then they do not know now that it will not be on the last day. They also know now that it will not be on the last day. Let p be 'The examination will be on the penultimate day' and l be 'The examination will be on the last day'. The assumptions may therefore be formalized as $K(p \supset \sim K{\sim}l)$ and $K{\sim}l$ respectively. We must distinguish between the pupils' reasoning and the reasoning of the theorist who propounds the paradox. The theorist reasons about the

[2] See Shaw 1958 and Kaplan and Montague 1960. Sorensen 1988: 298–310 criticizes the self-reference diagnosis.

[3] Sorensen 1988: 292–5 and 324–7 describes and criticizes attempts to assimilate the Surprise Examination to sorites paradoxes.

pupils' reasoning. The theorist infers from the premises K($p \supset$ ~K~l) and K~l the conclusion K~p. That inference is supposed to apply an assumption about the closure of the pupils' knowledge under their deductions to some reasoning that they perform. The pupils' conclusion is ~p. But if their premises are $p \supset$ ~K~l and ~l, then their conclusion does not follow from their premises, for if the examination were on the penultimate day and they knew that it would not be on the last day, their premises would be true and their conclusion false. To deduce ~p from the conditional premise $p \supset$ ~K~l, they need the extra premise K~l. But the theorist can apply the closure principle to that reasoning to infer that they know the conclusion ~p only on the assumption that they know the premises, and in particular that they know K~l. In other words, the theorist can conclude K~p only given the extra premise KK~l. But the theorist's original premise was only K~l. An instance of the KK principle is needed to bridge the gap. A corresponding instance is needed for each step of the backwards induction. Thus if the KK principle fails systematically for the reason explained in Chapter 5, the paradox of the Glimpse is dissolved.

Does that diagnosis of the Glimpse generalize to the Surprise Examination? The argument in the Surprise Examination can be reconstructed as using the KK principle; but it need not be.[4] A careful analysis shows that what the theorist really needs is the assumption that the pupils know on the first morning of term that they will know on the second morning that . . . they will know on the penultimate morning that they will know on the last morning the truth of the teacher's announcement. Similar iterations of knowledge are needed of the propositions that they will continue to be rational (deduce and come to know relevant consequences of what they know) and remember (know) on each morning before the examination that there has been no examination so far. The Glimpse too can be reconstructed with premises attributing the same number of iterations of knowledge to the pupils, but all concerning present knowledge, in place of the KK principle. The chief difference between the paradoxes is that in the Surprise Examination each iteration of knowledge involves a change in cognitive standpoint, whereas the Glimpse (like the KK principle) involves a fixed cognitive standpoint. Fortunately, the underlying objection to the KK principle in section 5.3 shows how both ascriptions of iterated knowledge can easily fail. The iteration of knowledge operators leads sooner or later to falsity through a process of erosion resulting from the need for margins for

[4] Harrison 1969 and McLelland and Chihara 1975 suggest rejection of the KK principle as a treatment of the Surprise Examination. It is criticized by Sorensen 1988: 312–20.

error. This applies just as much when the knowledge operators refer to different cognitive standpoints. If anything, it applies with more force: knowledge of future knowledge is usually less exact than knowledge of present knowledge, so wider margins for error are needed. The same point applies when the cognitive standpoints vary in other respects: I usually know less about your knowledge than about my own. Although section 5.5 revealed some unexpected complexities in the erosion process as margins for error are iterated, the general point remains that multiple iterations of knowledge are far harder to achieve than is usually recognized. Since the argument of both the Glimpse and the Surprise Examination make hidden assumptions of multiply iterated knowledge, the diagnosis from section 5.3 applies to both.

The diagnosis passes the test that it should allow the pupils to know the truth of the teacher's announcement. Since knowledge operators do not iterate automatically, it is consistent to suppose that the pupils know at the beginning of term that there will a surprise examination. If it were inconsistent, it would remain so if one ignored the distinction between what the pupils know at the beginning of term and what they know later: but it is provably consistent in the latter case. This result can be strengthened in various ways (see Appendix 2). This reinforces the earlier point that variation in cognitive standpoint is not essential to the problem, which can be raised and resolved even if it is axiomatic that the pupils never lose any knowledge, or if the argument is formulated with respect just to what they know at the start of term, as in #4. Since the pupils are more than one margin for error away from cases in which the teacher's announcement is false, they can know it to be true, but they are within a finite number of margins of such cases, so the number of iterations of knowledge they can have is limited.

Someone might object that the pupils can be idealized to a point where they do not need margins for error. Formally, 'the pupils know that' might be treated as an operator for provability from a fixed stock of assumptions in such a way that the KK principle holds. However, the paradox loses much of its interest under that idealization. In everyday cases, the pupils know that there will be a surprise examination, and this is what must be explained in the face of an apparent proof of its impossibility. In the idealized cases in which the KK principle holds, there is no presumption that the pupils can know that there is to be a surprise examination, so there is nothing to be explained. The paradox of the Glimpse disappears in the same way under the assumption that the pupils have perfect eyesight. Although variation in cognitive standpoint may sometimes allow the pupils to know that there will be a surprise examination even when the KK thesis holds, this does not apply in #1–#4, and it is probably not what makes it *obvious* that they can have the knowledge in ##1–##4 and ###1–###4.

The margins for error around the pupils' knowledge are part of the message of the Surprise Examination, not distracting noise. Some evidence for this is that the intuitive datum is not all-or-nothing but a matter of degree. Most people are simply confused by the one-day version of the paradox, 'There will be an examination today and you do not know it'. The more days are involved, the clearer it is that the pupils can know the truth of the teacher's announcement. We can explain this effect by supposing that a small difference between actual pupils and idealized ones is magnified by each stage of the reasoning until it is clearly visible. That difference is the margin for error.

The Surprise Examination is closely related to a number of paradoxes in decision theory.[5] The most prominent is Iterated Prisoner's Dilemma. Suppose that it is common knowledge for two agents that they are rational and face a series of ten thousand Prisoner's Dilemmas. Orthodox game theory 'proves' by a backward induction argument that they will never cooperate, for it is rational to cooperate at round *i* only if it is not then common knowledge that there is no round after *i* at which it is rational to cooperate. Defection dominates cooperation unless the latter makes future cooperation more probable. The result is highly unrealistic: ordinarily reasonable players are likely to cooperate for all but the last few rounds. Although such paradoxes will not be analysed in detail here, they all seem amenable to the present approach. In every case the argument assumes something like common knowledge of the agents' rationality: they are rational, both know that they are rational, both know that both know it, both know that both know that both know it, and so on. A standard game-theoretic backward induction argument invokes a further iteration of this knowledge for each round of the game. But any players we can envisage have less than perfectly accurate epistemic capacities, and know only if they leave a margin for error. The usual erosion may be expected to continue until falsity is reached after finitely many iterations of 'both know that'. Real agents may typically lack strict common knowledge of their rationality. The inexactness of their knowledge permits ordinarily reasonable players to cooperate in Iterated Prisoner's Dilemma.[6]

[5] Sorensen 1988: 333–70 discusses many examples. Rubinstein 1998: 165–74 briefly introduces some attempted solutions in the economics literature. See also Bovens 1997.

[6] See also Bicchieri 1989 and 1992, Bacharach 1992a and 1992b, and Rubinstein 1989. Pettit and Sugden 1989 deny the players common knowledge of their rationality, although not for the present reason; their treatment is related to Jackson's work on the Surprise Examination.

6.2 CONDITIONALLY UNEXPECTED EXAMINATIONS

The Surprise Examination and Iterated Prisoner's Dilemma differ slightly in structure. The pupils' argument assumes that there will be an examination on some day; if they did not assume this, a surprise examination could easily be held on the last day. The backward induction argument in Iterated Prisoner's Dilemma has no need to make the corresponding assumption: that the players will cooperate on some round. The purpose of this section is to eliminate the existential assumption from the Surprise Examination, and with it attempted solutions that treat it as essential.[7]

Consider first Mr Magoo. The existential assumption in the argument of section 5.1 is that the tree is some number of inches tall. However, this assumption is inessential to the argument. Suppose that Mr Magoo loses sight of the tree and hears tree-fellers at work. The tree is in danger of being felled. He does not know whether it still exists. If it does not, count every proposition of the form 'The tree is *i* inches tall' as simply false. Mr Magoo still knows that *if* the tree is *i*+1 inches tall then he does not know that it is not the case that the tree is *i* inches tall. Moreover, he knows that it is not the case that the tree is 0 inches tall, for he knows that if it still exists it is much taller than that. Given the KK principle and the background assumptions, the inductive argument in section 5.1 still 'shows' for any number *i* that Mr Magoo knows that it is not the case that the tree is *i* inches tall, and in particular that it is not the case that the tree is 666 inches tall. That is absurd. In fact, the tree still exists and is 666 inches tall. The existential assumption is irrelevant.

Now consider #1, the Glimpse. Suppose that the teacher is known to ring all examination dates and her family's birthdays on the calendar. The pupils do not know whether the ring they glimpsed is round an examination date or a birthday. As before, their glimpse is enough to see that one and only one date is ringed and that it is not very near the end of term but not to narrow it down much more than that. They still know now that, for all *i*, if there is an examination *i*+1 days from the end of term then they do not know now that there will be no examination *i* days from the end. They reason as before, concluding that there will be no examination. That is absurd, for we may stipulate that in fact the ring on the calendar does indicate an examination not very near the end of term. The existential assumption is irrelevant to the Glimpse.

[7] Sorensen 1988: 328–43a complicates his analysis of the Surprise Examination in order to generalize it to decision-theoretic paradoxes lacking an existential assumption. He has no version of the Surprise Examination itself without such an assumption.

Similar remarks apply to the other paradoxes in the array. Suffice it to eliminate the existential assumption from ###4, the Surprise Examination. Call an examination *conditionally expected* if on the morning of it the pupils know that if there is an examination at all, it will be that day. The teacher announces just that there will be no conditionally expected examination. The pupils argue from this that there will be no examination at all. If there has been no examination by the last morning, the pupils will certainly then know that if there is an examination at all it will be that day. For simplicity, we may assume that at most one examination can be held each term; the assumption can be eliminated from a more rigorous presentation of the argument. Thus a last-day examination would be conditionally expected, and so will not occur. The pupils will know this on the penultimate morning, so if there has been no examination by then, they will then know that if there is an examination at all, it will be that day. Thus a penultimate-day examination would be conditionally expected, and so will not occur. And so on. The pupils conclude that there will be no examination. That is absurd, for we may stipulate that in fact quite some time before the end of term there will be an examination that is not conditionally expected.

The Surprise Examination does not depend on the existential assumption. Indeed, it is a stronger paradox without it. The original argument was in effect a reductio ad absurdum (relative to background assumptions) of the supposition that the pupils know on the first morning that they know on the second morning that . . . they know on the last morning that there will be an unexpected examination. It would be obviously irrational to treat such an argument as justifying confidence that there will not be an examination. The new argument uses the supposition that the pupils know on the first morning that they know on the second morning that . . . they know on the last morning that there will be no conditionally expected examination to show that there will be no examination at all. This does not contradict the supposition; if there is no examination, the teacher's announcement is vacuously fulfilled. If the teacher and the pupils' memory and rationality are sufficiently reliable, why should they not treat the argument as justifying confidence that there will be no examination? It would be a quite inadequate response to remind the pupils that if there is an examination on the antepenultimate day (for example), then they do not know on the first morning that they know on the second morning that . . . they know on the last morning that there will be no conditionally expected examination. For why should they not treat the argument as justifying confidence that the antecedent of that conditional is false? Similarly, if I believe that my great-uncle died long after I was born on the grounds that I remember talking to him, it would be a quite inadequate response to remind me that if he died before I was born then I do not remember talking to him.

The point can be seen from the pupils' perspective. In the original version, the passage of day after day with no examination gradually undermines the pupils' justification for believing the teacher's announcement. It does no such thing in the new version. That there will be no examination looks increasingly likely: but this is the increasing likelihood of one way for the teacher's announcement to be fulfilled. As day after day passes with no examination, the pupils appear to remain justified in believing the teacher's announcement.

Consider, for instance, the pupils' situation on the penultimate morning, with no examination so far. The teacher's announcement now reduces to 'There will be no conditionally expected examination tomorrow and there will be no conditionally expected examination today'. Since any examination tomorrow will be conditionally expected for the trivial reason that no examination can be held after the end of term, the first conjunct boils down to 'There will be no examination tomorrow'. An examination today is conditionally expected just in case the pupils know that if there is an examination at all, it will be today, which in the circumstances is just for them to know that there will be no examination tomorrow; thus the second conjunct boils down to 'Either there will be no examination today or we do not know that there will be no examination tomorrow'. In the circumstances, the pupils know the truth of the teacher's announcement if and only if they know:

(!) There will be no examination tomorrow and either there will be no examination today or we do not know that there will be no examination tomorrow.

In symbols, (!) has the form $p \wedge (q \vee \sim Kp)$. Suppose that the pupils know (!). Thus they know its first conjunct; they know that there will be no examination tomorrow. Moreover, the second conjunct of (!) is true; either there will be no examination today or they do not know that there will be no examination tomorrow. By disjunctive syllogism, there will be no examination today. Thus if there is an examination today, the pupils do not know (!) and therefore do not know the truth of the teacher's announcement. However, (!) is not intrinsically unknowable. It follows logically from the straightforward conjunction 'There will be no examination tomorrow and there will be no examination today'. In the epistemic logic KT (see Appendix 2), $K(p \wedge (q \vee \sim Kp))$ entails $p \wedge q$ and is entailed by $K(p \wedge q)$; (!) might be called a contingent blindspot, or a contingently Moorean proposition. If the pupils knew the conjunction

'There will be no examination tomorrow and there will be no examination today', then they would know (!) and therefore the truth of the teacher's announcement. It is not illogical for the pupils to suppose themselves to know (!) in the way in which it would be illogical for them to suppose that they know 'There will be an examination today and we do not know that there will be an examination today' on the last morning of the original Surprise Examination.

Suppose that the very reliable teacher announces 'There will be no conditionally expected examination' at the beginning of a two-day 'term' during which there is in fact no examination. It is quite coherent to suppose that at the beginning of term the pupils know that they know the truth of the announcement, know that they know (!), and therefore know that there will be no examination. Nor is any incoherence introduced when the number of days in term and the number of iterations of knowledge is increased, provided that there is no examination. It just becomes ever more implausible that the pupils have the ever greater number of iterations of knowledge of the truth of the announcement needed to rule out its non-vacuous truth. This is precisely the erosion effect that margin for error principles predict.

Consider again Mr Magoo. Here the new version of the argument shows that if Mr Magoo has enough iterations of knowledge of (C), instances of (1_i), and some upper bound on the height of the tree (given its continued existence), then the tree does not exist and he knows it. But even if it does not exist, Mr Magoo cannot know that merely by reflecting on the limitations of his eyesight and ability to judge heights. So in the circumstances he cannot have that many iterations of knowledge of those propositions. In this case, the non-existence of the tree makes no more iterations of knowledge available to him than were already available when he knew that it existed. That is because the source of his knowledge is independent of the continued existence of the tree. In the Surprise Examination, by contrast, the source of the pupils' knowledge is the teacher who determines whether there will be an examination: if there is no examination as a result of her reliability, more iterations of knowledge may be available to the pupils. This difference is quite compatible with the assumption that margin for error principles govern both cases.

7
Sensitivity

7.1 PREVIEW

The argument of Chapters 4 and 5 connected knowledge and safety. If one knows, one could not easily have been wrong in a similar case. In that sense, one's belief is safely true. There is also a counterfactual notion of *sensitivity* to the truth, the simplest version of which requires that if the proposition were false, one would not believe it. Sensitivity does not entail safety. If *p* could not easily have been false, then one could not easily have falsely believed *p*; but that is consistent with the counterfactual that *if p had* been false, one would (or might) still have believed *p*. At first sight, that counterfactual looks like a reason for denying that one knows *p*. We might therefore conjecture that 'One knows *p*' entails 'If *p* were false, one would not believe *p*'; various more sophisticated conjectures can be made along similar lines. We might then wonder whether sensitivity should supplement or replace safety as a requirement on knowledge. This chapter argues that it should not.

The hypothesis that knowledge entails simple sensitivity has sceptical consequences. I believe that no evil demon is tricking me into believing that there are no evil demons. If an evil demon were tricking me into believing that there were no evil demons I would still believe that no evil demon was tricking me into believing that there were no evil demons. By the sensitivity hypothesis, I do not know that no evil demon is tricking me into believing that there are no evil demons. Is it bad faith to deny such sceptical consequences? Do they even confirm the general principle that knowledge requires sensitivity, since they allow it to explain the unwillingness which many feel to claim without qualification that they know that they are not in a sceptical scenario?

To simplify the story: famously, Robert Nozick once built 'If *p* were false, S would not believe *p*' as a conjunct into a conjunctive analysis of 'S knows *p*', thereby maintaining the entailment (Nozick 1981: 167–288). Problems dogged his account. In finely crafted recent work, Keith DeRose has argued for a pragmatic connection between knowledge attributions and such counterfactuals (DeRose 1995, 1996). Some of the problems that beset Nozick turn out to have analogues for DeRose.

This chapter will examine attempts to draw limited sceptical consequences from such counterfactual conditions on knowledge. To give those attempts their best chance, we must isolate them from the project of stating necessary and sufficient conditions for 'S knows *p*'. Attempts to provide jointly sufficient conditions for knowledge lead to conditions which are not individually necessary. For present purposes, what matters is only the more modest claim that a counterfactual condition is necessary for knowledge. As section 1.3 noted, that does not imply that knowledge is analysable. One upshot of the discussion will be that a sensitivity condition must keep the actual basis for the belief fixed in a way which undermines the sceptical argument from the sensitivity requirement.

7.2 COUNTERFACTUAL SENSITIVITY

Consider the principle:

(1) Necessarily, if S knows *p* then if *p* were false, S would not believe *p*.

If we write the counterfactual 'If *q* were true, *r* would be true' as $q \Box\!\!\rightarrow r$, we can symbolize (1) as $\Box(K_S p \supset (\sim p \Box\!\!\rightarrow \sim B_S p))$.[1] Let 'S sensitively believes *p*' abbreviate 'S believes *p*, and if *p* were false, S would not believe *p*'. Given that S knows *p* only if S believes *p* (as will be assumed throughout this chapter), (1) implies that S knows *p* only if S sensitively believes *p*. Any counterfactual conditional entails the corresponding material conditional, so in particular $\sim p \Box\!\!\rightarrow \sim B_S p$ entails $\sim p \supset \sim B_S p$. Thus 'If *p* were false, S would not believe *p*' is inconsistent with '*p* is false and S believes *p*', so S sensitively believes *p* only if *p* is not false. Given that *p* is true or false, S sensitively believes *p* only if *p* is true.

The most familiar semantics for the counterfactual conditional is that given by David Lewis (1973), on which $q \Box\!\!\rightarrow r$ is true at a possible world *w* if and only if either *q* is true at no possible world (the vacuous case) or, for at least one possible world *x*, *q* is true at *x* and *r* is true at every possible world at least as close in the relevant respects as *x* is to *w*. Nozick accepts something like this account when *q* is false at the world of evaluation *w*. But when *q* is true at *w*, Lewis's account implies that $q \Box\!\!\rightarrow r$ is true at *w* if and only if *r* is true at *w*, for the only world at

[1] For simplicity, the exposition follows the usual sloppy practice of not distinguishing between the proposition *p* and the proposition that *p* is true. In this context it is harmless.

least as close to *w* as *w* is *w* itself; thus $q \wedge r$ entails $q \Box\!\!\rightarrow r$. Nozick cannot accept that part of Lewis's account, for Nozick's analysis of knowledge includes a fourth conjunct (in addition to truth, belief, and the counterfactual linking falsity to unbelief) of the form 'If *p* were true, S would believe *p*' (Nozick 1981: 176). Lewis's account makes that conjunct redundant in the presence of the truth- and belief-conditions. In order to defeat some potential counterexamples to his analysis, Nozick interprets the fourth conjunct differently; its truth at a world *w* at which *p* is true requires S to believe *p* not just at *w* but at all worlds close to *w* at which *p* is true. On the corresponding modification of Lewis's account, if *q* is true at *w* then $q \Box\!\!\rightarrow r$ is true at *w* if and only if *r* is true at every world close to *w* at which *q* is true. This notion of closeness is not Lewis's comparative notion; it is more like the notion discussed in section 5.3. Indeed, on a Nozickian semantics for counterfactuals, one could express safety—the avoidance of false belief at close worlds—by contraposing the counterfactual in (1): 'If S were to believe *p*, *p* would be true'. Given that S does believe *p*, the Nozickian semantics makes that counterfactual true if and only if *p* is true at all close worlds at which S believes *p*. That counterfactual is not equivalent to the uncontraposed version, for $B_Sp \Box\!\!\rightarrow p$ can be true and $\sim p \Box\!\!\rightarrow \sim B_Sp$ false if *p* is true at every close world but S believes *p* at the closest (but not close) world at which *p* is false. Equally, $\sim p \Box\!\!\rightarrow \sim B_Sp$ can be true and $B_Sp \Box\!\!\rightarrow p$ false if S believes *p* at some close but not closest worlds at which *p* is false. Ernest Sosa has argued for something like the 'safety' conditional $B_Sp \Box\!\!\rightarrow p$ as a condition on knowledge (Sosa 1996, 2000). A more elaborate account on such lines would qualify 'S believes *p*' in the conditional to exclude cases in which S believes *p* on a quite different basis from the basis on which S believes *p* in the case in which S putatively knows *p*. Of course, the closeness account in Chapter 5 does not itself involve such a deviant account of counterfactuals; it is neutral on that issue, because it can be expressed in other terms. However, our present concern is with Nozick's original counterfactual $\sim p \Box\!\!\rightarrow \sim B_Sp$, for that is the one which differs most seriously from the closeness condition.

A slight variant of (1) in the same spirit says that 'S knows *p*' is incompatible with the opposite counterfactual 'If *p* were false, S would believe *p*'. It corresponds to replacing the counterfactual in (1) with the negation of the opposite counterfactual; in symbols: $\Box(K_Sp \supset \sim(\sim p \Box\!\!\rightarrow B_Sp))$. If 'might' is the dual of 'would', that is equivalent to the claim that, necessarily, if S knows *p* then if *p* were false, S might not believe *p*. On both Lewis's semantics and a Nozickian modification of it, the opposite counterfactuals $\sim p \Box\!\!\rightarrow B_Sp$ and $\sim p \Box\!\!\rightarrow \sim B_Sp$ are both false at the same world if S believes p at some but not all of the very closest worlds at which p is false. In such cases, S satisfies the 'might' condition for knowing but not the original 'would' condition in (1). Thus $\sim p \mathbin{\Box\!\!\rightarrow} \mathrm{B}_S p$ and $\sim p \mathbin{\Box\!\!\rightarrow} \sim \mathrm{B}_S p$ are not jointly exhaustive; they are almost mutually exclusive, for they are both true at the same world only if p is a necessary truth. In that case, S satisfies the 'would' condition for knowing in (1) vacuously, but automatically fails the 'might' condition. Thus if counterfactuals with impossible antecedents are vacuously true, the unrestricted generalization that 'S knows p' is incompatible with 'If p were false, S would believe p' has the absurd consequence that it is impossible to know necessary truths. In that respect, the original 'would' condition in (1) is more attractive. In any case, the difference between the 'would' and 'might' conditions does not matter for most of the subsequent discussion.

The appeal to possible worlds in the semantics of counterfactuals is controversial. Nevertheless, Lewis's semantics or a Nozickian modification of it will often be used in what follows, for it imposes a useful discipline on the discussion, and it informs the thinking of those who defend counterfactual conditions on knowledge. If another semantics of counterfactuals is preferred, the arguments below could be modified accordingly.

7.3 COUNTERFACTUALS AND SCEPTICISM

Let the *good case* be an ordinary situation in which things appear to be as they are, and the *bad case* be a sceptical scenario in which one falsely appears to oneself to be in the good case. In the good case, one believes truly that one is not in the bad case. In the bad case, one believes falsely that one is not in the bad case. Thus in the good case one does not sensitively believe that one is not in the bad case, for if that belief were false, one would be in the bad case and would still have that belief. Given (1), in the good case one does not know that one is not in the bad case. Right now, we do not know that we are not in a sceptical scenario.

That result depends on no specific theory of counterfactuals. We can individuate the bad case finely enough to determine what one believes in it, so that, necessarily, one is in the bad case only if one believes that one is not in the bad case. Then the antecedent of the counterfactual 'If one were in the bad case, one would not believe that one was in the bad case' is metaphysically possible, but not compossible with the consequent. On any reasonable account, including Lewis's and Nozick's, a counterfactual strictly implies that if the antecedent is possible, it is compossible with the consequent; in symbols, $\Box((q \mathrel{\Box\!\!\rightarrow} r) \supset (\Diamond q \supset \Diamond(q \wedge r)))$. On all such accounts, the counterfactual 'If one were in the bad case, one would not believe that one was in the bad case' is therefore false.

If one knows that which one deduces from what one knows, then by (1) one also fails to know in the good case any ordinary proposition *p* from which one deduces that one is not in the bad case, for if one knew *p* one would thereby know that one was not in the bad case, which by (1) one cannot. This holds even if in the good case one sensitively believes *p*; the problem is that one insensitively believes a consequence of *p*. If I were not typing I would be telephoning and I would not believe that I was typing; but from the proposition that I am typing I have deduced that I am not a brain in a vat falsely believing that I am typing. If I were a brain in a vat falsely believing that I was typing I would not believe that I was such a brain. By (1), I do not know that I am not a brain in a vat falsely believing that I am typing; if my knowledge is closed under my deductions, then I do not know that I am typing. Nozick rescues the possibility of knowing that I am typing, but not the possibility of knowing that I am not a brain in a vat falsely believing that I am typing, by embedding the consequent of (1) as a necessary condition in an analysis of knowledge on which the closure principle fails.[2] But accepting (1) commits one to neither Nozick's analysis nor non-closure; sceptics can consistently accept (1) while insisting on closure and rejecting the analysis. Accepting (1) commits one to a significant dose of scepticism. Accepting (1) with closure commits one to a massive dose of scepticism. Thus (1) commits one to the disjunction of non-closure and rampant scepticism.

DeRose has a subtler view, on which (1) is not universally true. He accepts a closure principle and regards many utterances of the form 'S knows *p*' in ordinary contexts as expressing truths. Nevertheless, he holds, such utterances generate contextual standards for the correct application of 'know' high enough to verify the corresponding instances of (1). Roughly speaking, 'S knows *p*' is true in a given context only if one avoids falsely believing *p* in the worlds relevant to that context; when that sentence is under consideration, the closest worlds in which *p* is false tend to become contextually relevant (as do worlds closer than they are to the original world). On DeRose's view, closure holds within contexts but not across them. Once the predicate 'know oneself to be in

[2] See also Dretske 1970. For relevant discussion of Nozick see Luper-Foy 1987, references therein, and Peacocke 1986: 127–52.

the bad case' is in play, the bad case becomes contextually relevant, standards for the correct application of 'know' rise accordingly, and sentences of the form 'S knows *p*' which expressed true propositions as uttered in ordinary contexts now express different and false propositions.

DeRose's account disagrees with Nozick's over the truth-values of utterances such as 'He knew that he was typing' said by you about me when sceptical possibilities have been raised to relevance in your context of utterance. For DeRose, you spoke falsely because my epistemic position was not good enough by the standards relevant to the context of your utterance. For Nozick, your utterance is true because I met his standards for knowledge and raising sceptical possibilities does not alter those standards; $\sim p \Box\!\!\rightarrow \sim B_S p$ can be true even if $\sim p \wedge B_S p$ is true at some contextually relevant possible worlds, provided that they are not the closest worlds at which $\sim p$ is true. For example, 'If I were not typing, I would not believe that I was typing' may be true even if in some contextually relevant sceptical possibility I falsely believe that I am typing, because in closer possible worlds in which I am not typing I believe that I am telephoning, not typing.

On a view intermediate between Nozick and DeRose, something like (1) is universally true but the truth of the counterfactual requires the truth of its consequent at all contextually relevant worlds at which the antecedent is true. For on a contextualist account of counterfactuals themselves, $q \Box\!\!\rightarrow r$ as uttered in a context c is true as evaluated with respect to a world w if and only if $\Box(p_c \supset (q \supset r))$ is true at w, where p_c is true at all and only the worlds relevant to c. This view makes $q \Box\!\!\rightarrow r$ equivalent to its contrapositive $\sim r \Box\!\!\rightarrow \sim q$ in any given context, although of course the contextual effects of *uttering* those counterfactuals may differ. In particular, Nozick's counterfactual $\sim p \Box\!\!\rightarrow \sim B_S p$ would be logically but not pragmatically equivalent to the safety conditional $B_S p \Box\!\!\rightarrow p$ in any given context on the contextualist view of counterfactuals.[3]

7.4 METHODS

Nozick himself gives a counterexample to (1):

> A grandmother sees her grandson is well when he comes to visit; but if he were sick or dead, others would tell her he was well to spare her upset. Yet this does

[3] Wright 1983 criticizes Nozick for neglecting context-dependence in counterfactuals.

not mean she doesn't know he is well (or at least ambulatory) when she sees him. (1981: 179)

The falsity of the grandmother's belief that her grandson is well in the bad case is consistent with her knowledge that he is well in the good case, because she does not believe that he is well by the same method in the two cases. In the good case, her belief is based on perception, in the bad case on testimony.

In response, Nozick reworks his analysis in terms of the technical notion of knowing via a method (or way of believing). We can extract a putative necessary condition for knowing via method M which stands to Nozicks's analysis of that notion as (1) stands to his original analysis of knowledge:

(2) Necessarily, if S knows *p* via method M, then if *p* were false and S were to use M to arrive at a belief whether *p*, S would not believe *p* via M.

If the grandson were not well and the grandmother were to use the method of visual inspection to arrive at a belief as to whether he was well, then she would not believe that he was well via that method. We should note that for Nozick, knowing *p* via some method or other is insufficient for knowing *p simpliciter* when one believes *p* via more than one method.

Nozick sometimes writes as though methods were reasonably general. But if they are, then (2) is false. To adapt an example from Goldman (1976: 779), I might know that I am seeing a dog when I look at a nearby dachshund in good light. In the circumstances, if I had not been seeing a dog, I would have been seeing a wolf, and would have falsely believed myself to be seeing a dog. My tendency to mistake wolves for dogs is consistent with my ability to recognize dachshunds as dogs, although it may impugn my ability to recognize alsatians as dogs. Let *p* be the proposition that I am seeing a dog. I believe *p* by the same general method in the two cases; I judge on the basis of sight, at the same distance and in the same light. If that is method M, then (2) is false.

Can we distinguish the methods by the specific difference in my visual evidence, as Goldman in effect does in his treatment of perceptual knowledge? Nozick sometimes mentions finely individuated methods, for instance, when he suggests inference from a specific proposition *q* as a method (1981: 189). Let E be the grandmother's particular evidence, and M the method of judging on the basis of E. In some possible situation, is her grandson unwell while she uses M to judge whether he is well? If not, then using M to judge whether he is well guarantees that he is well, perhaps because using M entails having E as part of one's evidence, E is part of one's evidence only if its constituent propositions are true, and those constituent propositions entail that her grandson is well. On that view, if her grandson is ill and she is presented with a perfect lookalike of him to keep her happy, then E is not part of her evidence, although she does not know the difference (see Chapters 8 and 9). That is not how Nozick wishes to individuate methods:

> The method used must be specified as having a certain generality if it is to play the appropriate role in subjunctives. This generality is set by the differences the person would notice; the methods are individuated from the inside. (1981: 233)

As Nozick individuates M, in some possible situation the grandson is unwell while the grandmother uses M to judge whether he is well. In that situation, she will presumably believe that he is well, because she is judging on the same internal evidence E as she actually has. Thus the counterfactual in (2) is false, and (2) still counts the grandmother as not knowing that her grandson is well, contrary to the intuition which prompted Nozick to invoke methods.

We can avoid the counterintuitive result by deleting the second conjunct from the antecedent of the counterfactual in (2):[4]

(3) Necessarily, if S knows *p* via method M then if *p* were false, S would not believe *p* via M.

Now, (2) entails (3) on most theories of counterfactuals, but (3) does not entail (2).[5] For (3) is consistent with the grandmother's knowing that her

[4] Luper-Foy 1984 has related arguments for this emendation.

[5] The counterfactuals in (2) and (3) are of the forms $(q \wedge (r \vee s)) \Box\!\!\rightarrow {\sim}r$ and $q \Box\!\!\rightarrow {\sim}r$ respectively. Call a world at which the antecedent and consequent of a counterfactual are both true, a *positive case* for the counterfactual, and a world at which the antecedent is true and the consequent false, a *negative case*. In general, we should expect the set of worlds at which a counterfactual is true to increase with its positive cases and decrease with its negative cases. More precisely, we should expect that if every positive case for the counterfactual $\mathbb{C}$ is a positive case for the counterfactual $\mathbb{C}^*$, and every negative case for $\mathbb{C}^*$ is a negative case for $\mathbb{C}$, then $\mathbb{C}^*$ is true at every world at which $\mathbb{C}$ is true. Lewis's semantics satisfies this constraint, for it makes a counterfactual true at *w* if and only if either it has no negative cases or, for at least one positive case *x*, no negative case is at least as close to *w* as *x* is. Since $(q \wedge (r \vee s)) \wedge {\sim}r$ entails $q \wedge {\sim}r$, every positive case for $(q \wedge (r \vee s)) \Box\!\!\rightarrow {\sim}r$ is a positive case for $q \Box\!\!\rightarrow {\sim}r$; since $q \wedge r$ entails $(q \wedge (r \vee s)) \wedge r$, every negative case for $q \Box\!\!\rightarrow {\sim}r$ is a negative case for $(q \wedge (r \vee s)) \Box\!\!\rightarrow {\sim}r$. Consequently, $(q \wedge (r \vee s)) \Box\!\!\rightarrow {\sim}r$ entails $q \Box\!\!\rightarrow {\sim}r$ and (2) entails (3) on Lewis's semantics and others meeting the constraint above. Matters are less straightforward on a Nozickian modification of Lewis's semantics. For suppose this: at *w*, *p* is false and S does not use M to arrive at a belief whether *p*; at the closest world *x* to *w* at which *p* is false and S uses M to arrive at a belief whether *p*, S does not believe *p* via M; at another world *y* close (but not as close as *x*) to M, *p* is false and S does believe *p* via M. Then the antecedent of the counterfactual in (2) is false at *w*, so we can evaluate the counterfactual there by Lewis's method, with the result that it is true at *w* thanks to *x*; but the antecedent of the counterfactual in (3) is true at *w*, so by the Nozickian modification the counterfactual is false

grandson is well via the method of judging on the basis of internally individuated perceptual evidence E. In the circumstances, if he were unwell her evidence would be distinguishably different from E. Likewise in the Goldman case: (3) is consistent with my knowing that I am seeing a dog via the method of judging on the basis of internally individuated perceptual evidence F.

Sometimes S does not know *p* via M and the counterfactual in (2) is false while that in (3) is true, so (2) but not (3) might explain why S does not know via M. For example, I cannot distinguish Tweedledee from Tweedledum by sight. I see them equally often. When I see either, I believe that Tweedledum is around. When I see neither, I form neither the belief that Tweedledum is around nor the belief that he is not. Even when I believe truly via the method of sight that Tweedledum is around, I do not know via that method that he is around. Suppose that on this occasion if Tweedledum had been absent, so would Tweedledee have been. Since I would have formed no belief either way, the counterfactual in (3) is true for 'Tweedledum is around' in place of '*p*'. By contrast, the counterfactual in (2) is false, for if Tweedledum had not been around and I had formed a belief one way or the other on the basis of sight, it would have been by seeing Tweedledee, and I would falsely have believed that Tweedledum was around. Of course, such cases do not show that (3) is false, just that it cannot explain every failure to know. We are not assessing (3) as a candidate for use in a supposedly necessary and sufficient condition for knowledge, such as Nozick sought.

Unlike (1), (3) does not automatically generate sceptical consequences. Everything depends on the individuation of methods. Suppose that in the good case one believes via M that one is not in the bad case. Then (3) forbids this true belief to constitute knowledge that one is not in the bad case via M only if in the bad case one believes *via M* that one is not in the bad case. For example, if one's belief that one is not in the bad case derives in the good case but not in the bad case from seeing one's body, why should that not constitute a difference in the method used? Nozick and others will object that such a difference does not count because it is inaccessible to the subject. But that externally individuated methods are inaccessible to the subject is far from obvious. The anti-sceptic will insist that in the good case one does know that one

at *w* thanks to *y*. Consequently, on the envisaged modification of Lewis's semantics, the counterfactual does not satisfy the constraint formulated above. That is probably a reason for a further modification to bring the semantics into line with the constraint. It does not make much practical difference, since the divergence occurs only at worlds at which *p* is false; since knowing via method M is factive, S anyway fails to know *p* via M at such worlds, so the difference in the counterfactuals does not imply that (3) can be false if (2) is true.

is seeing one's body, even if in the bad case one falsely believes that one is seeing one's body. To assume in deriving sceptical consequences from (3) that in the good case one does not know that one is seeing one's body would beg the question. To serve its dialectical purpose, the accessibility claim should be that the method M must be so individuated that in *every* possible case (not just every case in which one knows via M) one is in a position to know whether one is using M. But the argument of Chapter 4 suggests that *no* principle of individuation yields that result. However methods are individuated, one is not always in a position to know whether one is using M.

If methods are individuated internally, so that whether one is using method M supervenes on the physical state of one's brain, then (3) will indeed have some sceptical consequences. But why should one accept (3) on those terms? The internal individuation of methods appears gerrymandered precisely to make trouble for our claims to knowledge of the external world. Moreover, (3) is implausible in some examples when methods are individuated internally. My knowing by sight in the ordinary way that a mountain stands here seems compatible with the assumption that if no mountain had stood here, a bizarre chain of circumstances would have occurred as a result of which I would have hallucinated a mountain of just this appearance. That type of hallucination occurs only in worlds very unlike the actual world, we may suppose, and the mechanism that produces it is absent from the actual world. I actually satisfy (3) for knowing by sight many other things about my present environment, including that there is an *icy* mountain here; my eyesight is actually working perfectly and I have every ordinary reason to believe that it is. To block the unwarranted consequence of (3) that I do not know that a mountain stands here, one must individuate methods externally rather than internally.[6] The next two chapters develop the argument for the external individuation of methods and evidence.

7.5 CONTEXTUALIST SENSITIVITY

DeRose avoids internalist gerrymandering. He responds to Nozick's grandmother without appealing to internally individuated methods.

DeRose first suggests that we might defend the unmodified (1) in Nozick's example by giving heavy weight to similarity in method of

[6] See Luper-Foy 1984 and Shatz 1987 for the external individuation of methods in relation to Nozick. For other criticisms of (3) see Vogel 1987.

belief formation in the overall similarity measure between possible worlds which we use to evaluate the counterfactual. In determining what would be true if *p* were false we concentrate on worlds in which S believes via the same method (1995: 20–1). The grandmother knows that her grandson is well; if he were ill she would not believe that he was well, because she could see that he was not well. One might also try to analyse the example of the counterfactually hallucinated mountain by assessing worlds at which there is no hallucination and no mountain as more similar than worlds in which there is a hallucination and no mountain to the actual world in which there is a mountain and no hallucination, on the grounds that the introduction of a highly unusual causal mechanism makes a highly significant difference to the method of belief formation. Thus 'If no mountain stood here, I would not believe that a mountain stood here' might be forced to come out true. But loading the similarity relation is less likely to help with Goldman's dog. If seeing the grandson well and seeing him ill make for relevantly similar methods of belief formation, why not seeing a dachshund and seeing a wolf?

DeRose does not claim that weighting the similarity relation can handle every example. For instance, if *p* is the proposition that I don't falsely believe that I have hands, then the counterfactual in (1) is false whatever reasonable similarity relation is used. If I falsely believed that I had hands, I would believe that I had hands and believe the logical consequence that I didn't falsely believe that I had hands. Intuitively, however, this example hardly threatens my knowledge that I don't falsely believe that I have hands; (1) is false in this instance. The same goes for my knowledge that I am not an intelligent dog who is always incorrectly thinking that it has hands (1995: 22).

DeRose proposes a rough qualification of (1):

> We *don't*... judge ourselves ignorant of *p* where ~*p* implies something we take ourselves to know to be false, without providing an explanation of how we came to falsely believe this thing we think we know. Thus, *I falsely believe that I have hands* implies that I don't have hands. Since I do take myself to know that I have hands (*this* belief isn't insensitive), and since the above italicized proposition doesn't explain how I went wrong with respect to my having hands, I'll judge that I do know that proposition to be false. (1995: 23; symbolism modified)

The passage suggests that the consequence of ~*p* we take ourselves to know to be false ('I don't have hands') is falsely believed to be false if *p* is false, and that ~*p* does not explain why. From a third-personal perspective, we presumably judge S to know *p* in these cases only if S believes *p* at least in part *because* ~*p* has the consequence which we take

S to know to be false. S does not know *p* if S believes *p* only on different and bad grounds. This reference to the grounds on which S believes is a step in the direction of Nozick's methods, but not far enough to block the account's partially sceptical consequences. The negation of the proposition that I am not a brain in a vat induced to appear to itself to have hands *does* explain how I could falsely believe that I have hands, because that sceptical scenario has enough detail to be explanatory.[7]

DeRose's account as just stated is insufficiently general. For example, let α be a case in which I am climbing a mountain and appear to myself to be climbing a mountain; things appear to me as they are. In case β, I am a brain in a vat, but things appear to me generally just as they do in α. If I were in β, I should believe that I was not in β. If *p* is the proposition that I am not in β, then I do not sensitively believe *p*. Nevertheless, in my present case, I know that I am not in β, because I am in neither α nor β; things do not appear to me at all as they would in β. I appear to myself to be sitting in front of a computer screen in my office. I do not appear to myself to be climbing a mountain, as I would in β.[8] No matter what my situation, I cannot sensitively believe *p*. Thus (1) is false. Moreover, my false beliefs when I am in β are just as explicable as in other sceptical scenarios, while the example does not involve false beliefs when I am in either α or my actual case. Thus the example does not involve unexplained false belief at all, contrary to the apparent suggestion in the quoted passage.

A small modification of DeRose's suggestion does handle the example. For let *q* be the proposition that I do not appear to myself to be climbing a mountain. Thus *q* entails *p* (that I am not in β), so ~*p* entails ~*q*. If I am in β, then I appear to myself to be climbing a mountain. I take myself to know *q*; it is just a belief about how things appear to me, and I am mistaken about *q* in none of the relevant cases. I sensitively believe *q*, for if I appeared to myself to be climbing a mountain I would not believe that I did not appear to myself to be climbing a mountain. Moreover, in β I do not falsely believe *q*; I believe ~*q*, truly. Thus ~*p* does not explain how I could falsely believe *q*. One might therefore

[7] The quoted passage assumes that we can treat my belief that I have hands as sensitive and my belief that I don't falsely believe that I have hands as insensitive in the same context. This implies the non-transitivity of the counterfactual conditional in a fixed context: 'If I falsely believed that I had hands, I wouldn't have hands' and 'If I didn't have hands, I wouldn't believe that I had hands' are true but 'If I falsely believed that I had hands, I wouldn't believe that I had hands' is false. Contextualist theories of counterfactuals seek to avoid this combination; see Wright 1983.

[8] Stephen Schiffer raises a similar problem for the sceptical hypothesis that I am a BIV′, which is just like a brain in a vat except for lacking auditory sensations (1996: 331).

modify DeRose's suggestion thus: when we judge (1) false, we do so because S sensitively believes a proposition *q* which entails *p*, and ~*p* does not explain how S could falsely believe *q*. When S does sensitively believe *p*, *p* is itself such a proposition *q*, granted that ~*p* does not explain how one could falsely believe *p* when one would not believe *p* if *p* were false. We might therefore modify (1) thus:

> (4) Necessarily, if S knows *p* then, for some proposition *q*: *q* entails *p*, S sensitively believes *q*, and ~*p* does not explain how S could falsely believe *q*.

Further modifications could be made. We might require that S believes *p* *because* S believes *q*. We might allow the link between *q* and *p* to be looser than entailment. The discussion below will not depend on these details. The contextualist motivation for replacing (1) by (4) is that when we evaluate 'S knows *p*' we need not consider cases in which S falsely believes *q*, because we have no contextually relevant explanation of how S could do so. The quoted passage suggests in addition the converse of something like (4), since DeRose speaks of circumstances where '[w]e *don't* . . . judge ourselves ignorant of *p*', but our present concern is not with sufficient conditions for knowledge. Of course, the contextualist's interest in (4) not as a universal principle but as a form instances of which tend to come out true when the corresponding instance of 'S knows *p*' is uttered.

How does (4) handle Goldman's dog? The obvious proposal is that although my belief that I am seeing a dog is insensitive, it is derived from a sensitive belief that I am seeing a dachshund. The hypothesis that I am not seeing a dog does not explain how I could falsely believe that I am seeing a dachshund. This proposal raises the question: even if I do have the intermediate belief that I am seeing a dachshund, why should it be sensitive? Perhaps I tend to mistake another similar breed of dog for dachshunds; I never mistake anything but dogs for dachshunds. Yet, when I see a dachshund, I still know that I am seeing a dog. Knowledge seems to be compatible with very widespread slight insensitivity of this kind. We can confirm the suspicion with another example.

I tend slightly to underestimate the distances I see. When I see a distance of twenty-one metres I judge it to be less than twenty metres, although when I see a distance of twenty-three metres I do not judge it to be less than twenty metres. This may mean that when I see a distance of nineteen metres and correctly judge it to be less than twenty metres, I do not know it to be less than twenty metres. It surely does not mean that when I see a distance of one metre and correctly judge it to be less than twenty metres, I do not know it to be less than twenty metres. A distance of one metre and a distance of twenty-one metres look quite different to me. My unreliability in answering the question 'Is it less than twenty metres?' when presented with one distance does not imply unreliability in answering it when presented with the other. Suppose that a mark on the side of a ship is one metre above the waterline, in circumstances in which if it had not been less than twenty metres above the waterline it might have been less than twenty-one metres above the waterline.[9] I judge by sight whether the mark is less than twenty metres above the waterline. Let *p* be the proposition that the mark is less than twenty metres above the waterline. If *p* had been false, I might still have believed *p*. I believe *p* insensitively. Surely I can still know *p*, because I believe *p* on quite different evidence from that on which I would have believed it had it been false. By (4), I must sensitively believe a proposition *q* entailing *p*, where ~*p* does not explain how I could falsely believe *q*. What is *q*? An obvious candidate is the proposition that the mark is less than two metres above the waterline. For that entails that it is less than twenty metres above the waterline; the hypothesis that the mark is not less than twenty metres above the waterline does not explain how I could falsely believe that it is less than two metres above the waterline. But I can know *p* even if I do not derive *p* from anything like the proposition that the mark is less than two metres above the waterline; *p* may be the only proposition which I entertain about the distance in metres. Even if I do believe that the mark is less than two metres above the waterline, that belief too may be insensitive. I have a general tendency to underestimate distances; if the mark were not less than two metres above the waterline, it might be only slightly more than two metres above, and I would still judge it to be less than two metres above the waterline. Even so, I can know that it is less than twenty metres above. Alternatively, I might derive *p* from the premises that the distance has a look L and that only distances of less than twenty metres have L. But do I believe the latter premise sensitively? Very likely not. If distances of slightly more than twenty metres had L, I would or might still believe that only distances of less than twenty metres have L. But I still know that the mark is less than twenty metres above the waterline. Thus (4) is vulnerable to widespread slight insensitivity of a kind compatible with knowledge. Appendix 3 gives a formal model to show how the slightest systematic inaccuracy can make one almost totally insensitive.

[9] 'Might' counterfactuals are here used as the negations of the opposite 'would' counterfactuals. If sensitivity were weakened to require only the falsity of 'If *p* were false, S would believe *p*', we could posit circumstances in which if the distance had been not less than twenty metres it would have been less than twenty-one metres.

A different proposal is to take *degree of belief* into account. The idea is that if the mark had been slightly more than twenty metres above the waterline, I would still have believed that it was less than twenty metres above the waterline, although with less confidence than I believe it when the distance is only one metre. But what if I am not like that? Suppose that once I form a belief in a marginal case, I stick to it; perhaps a macho mechanism causes me to feel an aggressive confidence in it even greater than I feel in non-marginal cases. Regrettable though that may be, when the distance is one metre it does not prevent me from knowing that it is less than twenty metres. Creatures whose beliefs are all or nothing in degree can have such knowledge.

The problem arises however small the inaccuracy is, if non-zero. Even if the only distances which I falsely believe to be less than *n* metres are less than a millimetre more than *n* metres, that belief is insensitive when, if the distance had been greater, it might have been less than a millimetre greater.

Naturally, individual examples do not refute the hypothesis that *most* ordinary cases conform to (4), or even to (1). DeRose prudently avoids advancing such principles as exceptionless generalizations; context-dependence is an unruly phenomenon. Nevertheless, he does not dismiss recalcitrant cases as statistically insignificant; he accepts the responsibility to explain them, as his willingness to replace (1) by something like (4) shows. It is quite unclear how to explain the counterexamples to (4) within a counterfactual framework without appeal to finely individuated methods, as in (2) or (3).[10]

7.6 SENSITIVITY AND BROAD CONTENT

The use of something like condition (4) to explain the appeal of scepticism faces a further problem, from the external individuation of content.

Hilary Putnam (1981) famously denied that a brain in a vat believes falsely that it is not a brain in a vat, arguing that it lacks the kind of causal connections to brains and vats needed to refer to them. If I were a brain in a vat, I might think 'I am not a brain in a vat', but would not thereby express the proposition that I am not a brain in a vat. If in the

[10] Both (4) and the proposal about degrees of belief are based on suggestions made by Keith DeRose, although he is not responsible for the details of the present formulations.

bad case one lacks the causal or conceptual resources to grasp the proposition that one is not in the bad case, then whenever one believes that one is not in the bad case one does so both truly and sensitively.

Putnam's treatment of scepticism looks insufficiently general. If I have only recently been envatted, I retain enough causal connection with my previous environment to formulate the proposition that I have not been envatted (Smith 1984, Wright 1992a: 86; DeRose 1995: 1). Nevertheless, more limited forms of content externalism threaten any attempt to extract sceptical consequences from (4).

Let the bad case be a moderate sceptical scenario in which content externalism does not prevent me from grasping my predicament. Let *p* be the proposition that I am not in the bad case. Then, in the bad case, I grasp *p*; I falsely believe *p*. Therefore, in the corresponding good case, in which things are as they appear to be in the bad case, I believe *p* truly but insensitively. Does (4) imply that I do not know *p*? In the good case I may believe sensitively a proposition *q* from which I deduce *p*, and content externalism may prevent me from grasping *q* in the bad case. For instance, let *q* be the proposition that I see this pen; *q* entails *p*. In the bad case, I cannot see this pen and have never had any contact with it, however indirect. I am not thinking about *this* pen, even if I am thinking about pens in general, and perhaps about a particular pen-image. Although I know the linguistic meaning of the words 'I see this pen', I cannot grasp the proposition that I see *this* pen. It is not expressed by that sentence in this context. Since I cannot believe *q* when *p* is false, ~*p* does not explain how I could falsely believe *q*. In the good case, I believe sensitively that I see this pen. If I did not see it, I would presumably be in some other ordinary situation and would not believe that I saw it. Even if I would or might be in the bad case if *q* were false, I would still not believe *q*; loading the similarity relation cannot help DeRose here. In the good case, I deduce that, since I see this pen, I am not in the bad case, and believe that I am not in the bad case. In the bad case, I still have the premise token 'I see this pen', and go through a process superficially like deduction, but my premise token lacks content. Thus the consequent in (4) is true for the specified values of '*p*' and '*q*'. For all (4) implies, 'I know that I am not in the bad case' is true after all. But that is a paradigm of the kind of knowledge claim which the contextualist account was supposed to falsify. Contextualists argue for contextualism by citing its ability to explain the appeal of scepticism, because it predicts the truth of many sceptical utterances. Those predictions follow from (1); they do not follow from (4). Indeed, the opposite predictions follow if content externalism and the converse of (4) are added to the theory.

Nozick might handle the example by insisting that I believe *p* via the same internally individuated method M in the two cases, whether or not I grasp *q*. Thus in the good case, although I truly believe *p* via M, if *p* had been false I would still have believed *p* via M; by (3), even in the good case I do not know *p*. But that sceptical result depends on a principle for individuating methods which was seen to be problematic in section 7.3.

Could DeRose handle the problem by modifying (4) to speak of quasi-propositions rather than propositions, where quasi-propositions are like propositions except for being individuated narrowly? That is easier said than done, for we need to be told *which* narrow criterion is to individuate quasi-propositions; which internal differences are consistent with thinking the same quasi-proposition? Furthermore, the appeal to quasi-propositions looks suspiciously ad hoc. It needs some independent motivation if it is not to be a mere device for rigging (4) in favour of the sceptic. Such a regression to internalist modes of thought also looks foreign to the otherwise externalist spirit of DeRose's account. The argument of the next chapter further undermines those modes of thought. In any case, we must recall that (4) remains defeated by the examples in section 7.5 of knowledge combined with systematic slight insensitivity.

8
Scepticism

8.1 PLAN

Rational thinkers respect their evidence. Properly understood, that is a platitude. But how can one respect one's evidence unless one knows what it is? So must not rational thinkers know what their evidence is? If so, then for rational subjects the condition that one has such-and-such evidence should be non-trivial yet luminous. But how can it be, given the anti-luminosity argument of section 4.3?

The assumption that rational thinkers know (or are in a position to know) what their evidence is has implications for sceptical arguments. Non-sceptics postulate a special asymmetry between the good and bad cases in a sceptical argument (section 8.2). Sceptics try to undermine the asymmetry by claiming that the subject has exactly the same evidence in the two cases, but this claim is not obvious (section 8.3). We can argue from the premise that rational thinkers know what their evidence is to the conclusion that their evidence is the same in the two cases (section 8.4). That conclusion forces one into a phenomenal conception of evidence (section 8.5). But the premise that rational thinkers know what their evidence is leads by a parallel argument to a clearly false conclusion (section 8.6). This is another variation on the arguments of sections 4.3 and 5.1. Rational thinkers are not always in a position to know what their evidence is; they are not always in a position to know what rationality requires of them (section 8.7). These conclusions generalize to sceptical arguments in which the sceptic does not claim sameness of evidence between the good and bad cases (section 8.8). One upshot is that sceptical arguments may go wrong by assuming too *much* knowledge; by sacrificing something in self-knowledge to the sceptic, we stand to gain far more in knowledge of the world.

8.2 SCEPTICISM AND THE NON-SYMMETRY OF EPISTEMIC ACCESSIBILITY

For simplicity, we can treat the sceptic as a generic figure, without attempting to track the protean variety of sceptical argument. Scepticism is a disease individuated by its symptoms (such as immoderate protestations of ignorance); we should therefore not assume that it can be caused in only one way. The present aim is to identify one main such way, not to eliminate the disease entirely.

For the sake of argument, let us assume that the constraints of the content externalism discussed in sections 2.2 and 3.2 are consistent with grasping the relevant propositions in sceptical scenarios. A recently envatted brain can still think about the external world. Even for such a brain, the assumption remains problematic as applied to propositions expressed by means of perceptual demonstratives (see also section 7.6). Suppose, for example, that I am looking at a cloud and think that that cloud is dark. A brain envatted before the advent of that cloud, with experience in some sense indistinguishable from mine, does not think that *that* cloud is dark, although it may think the words 'That cloud is dark' and be in no position to know that it does not thereby express a singular proposition concerning some cloud to the effect that it is dark. Similar issues arise for much less extravagant sceptical scenarios, involving mere hallucinations and the like. We assume for the sake of argument, perhaps over-generously, that the sceptic has some way of absorbing such implications of content externalism.

The sceptic compares a good case with a bad one. In the good case, things appear generally as they ordinarily do, and are that way; one believes some proposition *p* (for example, that one has hands), and *p* is true; by ordinary standards, one knows *p*. In the bad case, things still appear generally as they ordinarily do, but are some other way; one still believes *p*, but *p* is false; by any standards, one fails to know *p*, for only true propositions are known. As far as externalism permits, things appear to one in exactly the same way in the good and bad cases. The sceptic argues that because one believes *p* falsely in the bad case, one does not know *p* (even though *p* is true) in the good case. Let us postpone asking why the sceptic should think that false belief in one case precludes knowledge in the other, and consider the bad case.

Uncontroversially, if one is in the bad case then one does not know that one is not in the good case. Even if one pessimistically believes that one is not in the good case, one's true belief does not constitute knowledge; one has no reason to suppose that the appearances are misleading to that extent. More generally, it is consistent with everything one knows in the bad case that one is in the good case. For even if in the bad case one believes some true propositions which entail that (contrary to the appearances) one is not in the good case, those true beliefs do not all constitute knowledge. Part of the badness of the bad case is that one cannot know just how bad one's case is.

For the sceptic, the two cases are symmetrical: just as it is consistent with everything one knows in the bad case that one is in the good case, so it is consistent with everything one knows in the good case that one is in the bad case. One simply cannot tell which case one is in. For the sceptic's opponent, the two cases are not symmetrical: although it is consistent with everything one knows in the bad case that one is in the good case, it is not consistent with everything one knows in the good case that one is in the bad case. For in the good case, according to the sceptic's opponent, one knows *p* (for example, that one has hands), and also (by description of the bad case) that if one is in the bad case then *p* is false. These three propositions are jointly inconsistent:

(a) One is in the bad case.

(b) If one is in the bad case then *p* is false.

(c) *p*.

That argument does not assume that one knows that which is a logical consequence of what one knows, for the anti-sceptic's conclusion was merely that it is inconsistent with what one knows in the good case that one is in the bad case, not that one knows in the good case that one is not in the bad case. Although the anti-sceptic may hold that in the good case one also knows that one is not in the bad case, the asymmetry does not require that further knowledge claim.

We can state the asymmetry in the terminology of epistemic logic (see also section 10.4). A case β is said to be *epistemically accessible* from a case α if and only if everything which one knows in α is true in β. Then, according to the anti-sceptic, although the good case is epistemically accessible from the bad case, the bad case is not epistemically accessible from the good case.

Some refinements may be needed to handle the issues raised by the broad content of indexical expressions. As uttered in any case α, the sentence 'This case obtains' expresses a content true in α and in no other case. Perhaps one can know that content in α without knowing everything about α; we might allow cases other than α to be epistemically accessible from α, on the grounds that 'This case obtains' expresses (different) true contents in them. This complication does not affect the main arguments to come.

As is well known, asymmetries of epistemic accessibility yield counterexamples to the epistemic version of the 'Brouwersche' thesis in modal logic, the principle that if *p* is false then one knows that one does not know *p* ($\sim p \supset \mathrm{K}{\sim}\mathrm{K}p$; 'K' for 'one knows that'), and consequently to the epistemic version of the S5 thesis, the principle that if one does not know *p* then one knows that one does not know *p* ($\sim\mathrm{K}p \supset \mathrm{K}{\sim}\mathrm{K}p$).

The latter principle entails the former because knowledge is factive ($Kp \supset p$ holds). Like epistemic S4 (the KK principle), epistemic S5 embodies a luminosity claim. But the failure of the epistemic S5 principle on non-sceptical assumptions was already noted in section 1.2, independently of general anti-luminosity arguments. For in the bad case, *p* is false and one does not know *p*, but one does not know that one does not know *p*. If one knew in the bad case that one did not know *p*, then according to the sceptic's opponent it would not be consistent with everything one knew in the bad case that one was in the good case, since these three propositions are jointly inconsistent:

(d) One is in the good case.

(e) If one is in the good case then one knows *p*.

(f) One does not know *p*.

According to the sceptic's opponent, one can know (e) even in the bad case by description of the good case and one's appreciation that it meets the conditions for one to know *p*. The failure to know that one fails to know is characteristic of the bad case. Although the sceptic will try to argue that the postulated asymmetry between the two cases is ultimately unstable, there is at least no immediate incoherence.[1]

A common means of slurring over the epistemic asymmetry is to speak of the two cases as indiscriminable. Surely, if *x* is indiscriminable from *y* then *y* is indiscriminable from *x*. But even indiscriminability embodies a concealed asymmetry. For one may be able to discriminate between *x* and *y* when they are presented in one way and not when they are presented in another (Williamson 1990a: 14-20). A case can be presented in two relevant ways. When one is in a case, one can present it indexically to oneself, as 'my present case'. Alternatively, whether one is in a case or not, one can present it descriptively to oneself, for example, the good case as 'the good case' and the bad case as 'the bad case'. Since we have two cases and two modes of presentation of each of them, we have the four possibilities in Table 1 to consider.

Possibility II does not arise, because a case can be presented indexically as 'my present case' only if one is in it; since one cannot be in both the good and bad cases simultaneously, one cannot be faced with the task of discriminating between them, each presented indexically as 'my

[1] For epistemic asymmetry in relation to scepticism see Williams 1978: 310–13, although Williams is confident of an asymmetry only when death, drugs, sleep, or the like incapacitate the subject from thinking rationally. Humberstone 1988 has a subtle discussion of obstacles to asymmetry. For further sources of epistemic asymmetry, see section 10.4 and Appendix 5.

TABLE 1. *Presentation of Cases*

Possibility	Presentation of good case	Presentation of bad case
II	Indexical: 'my case'	Indexical: 'my case'
ID	Indexical: 'my case'	Descriptive: 'the bad case'
DI	Descriptive: 'the good case'	Indexical: 'my case'
DD	Descriptive: 'the good case'	Descriptive: 'the bad case'

present case'. DD discrimination is trivial, for one is merely required to discriminate conceptually between them presented as 'the good case' and 'the bad case', with no need to discover which case one is in. The interesting possibilities are ID and DI. Sceptics and anti-sceptics agree that in the bad case one cannot discriminate the bad case, presented indexically as 'my present case', from the good case, presented descriptively as 'the good case'. Thus it is uncontentious that the cases are DI indiscriminable. The issue is whether they are ID indiscriminable. Indiscriminability is symmetric in the sense that if x presented under mode M is indiscriminable from y presented under mode N, then y presented under mode N is indiscriminable from x presented under mode M, but it obviously does not follow that x presented under mode N is indiscriminable from y presented under mode M. DI indiscriminability does not imply ID indiscriminability. The anti-sceptic claims that in the good case one *can* discriminate the good case, presented indexically as 'my present case', from the bad case, presented descriptively as 'the bad case', for that is just to know in the good case that one is not in the bad case. The sceptic claims that one cannot make that discrimination, but since that is in effect to claim that in the good case one cannot know that one is not in the bad case, ID indiscriminability is tantamount to the sceptic's conclusion. The sceptic cannot use it as a premise without begging the question.

In a more complex version of the argument, the sceptic may postulate a subject whose case oscillates over time between the good case and the bad case. Such a subject may indeed be incapable of discriminating between the good case, presented indexically as 'my present case', and the bad case, presented indexically as 'my case five minutes ago', and therefore lack the relevant knowledge. It does not follow that one lacks that knowledge even if one's case is not in fact oscillating, or in danger of doing so. Thus the oscillation example does not achieve the sceptic's purpose. Alternatively, the sceptic may prefer to work with identity of appearance rather than with indiscriminability. The ultimate uselessness of such an appeal will emerge in the course of the argument below.

8.3 DIFFERENCE OF EVIDENCE IN GOOD AND BAD CASES

The sceptic typically insists that one has exactly the same evidence in the two cases. Therefore, since one believes *p* with that evidence in the bad case, believing *p* with the evidence one has in the good case is insufficient for the truth of *p*. If the sceptic allowed that one had different evidence in the two cases, false belief in the bad case would be a far less pressing threat to knowledge in the good case: the possibility of falsely believing *p* on the basis of bad evidence is quite compatible with the possibility of knowing *p* on the basis of good evidence. Scepticism about the external world has more intuitive force than scepticism about one's own sensations because we do not usually envisage beliefs about one's own sensations as based on evidence insufficient for their truth.

The sceptic cannot simply stipulate that one has the same evidence in the good and bad cases. For the notion of evidence will serve the sceptic's purposes only if it has non-trivial connections with other epistemic notions, such as the notion of knowledge. Some externalists about evidence (although not all) will argue that those connections force a difference in evidence between the two cases. If the sceptic tries to stipulate that the bad case is a case in which one falsely believes *p* while having the same evidence as one has in a case in which by externalist standards one knows *p*, those externalists will reply that, so defined, the bad case is impossible, and the sceptic's argument does not get off the ground. Rather, the sceptic should define the bad case in less contested terms, so that its possibility is agreed, and then *argue* for the lemma that one has the same evidence in it as in the good case. Many contemporary non-sceptics accept that lemma in the sceptic's overall argument. They concede that when we have empirical knowledge, we could have had false belief in the same proposition with exactly the same evidence. Many hold that, at least in some contexts, the bad case is in some sense irrelevant to the attribution of knowledge in the good case.[2] For present purposes, what matters is simply the claim that one has the same evidence in the two cases. How can that claim be supported?

A natural argument is by reductio ad absurdum. Suppose that one has different evidence in the two cases. Then one can deduce in the bad case that one is not in the good case, because one's evidence is not what

[2] Lewis 1996 gives a recent account of this kind in which sameness of evidence plays a central role. McDowell 1982 denies that the evidence is the same. For the relevant alternatives approach generally see Goldman 1976, Stine 1976, Dretske 1981b, and Cohen 1988.

it would be if one were in the good case. But even the sceptic's opponent agrees that it is consistent with everything one knows in the bad case that one is in the good case. Therefore, one has the same evidence in the two cases.

The argument assumes that in the bad case one knows what one's evidence is, otherwise one would lack a premise for the deduction. Now, surely one can be rational even in the bad case; misleading evidence sometimes makes false beliefs rational. So one can know what one's evidence is, granted the assumption that rational thinkers are in a position to know what their evidence is. The appeal of that assumption is by no means limited to sceptics; after all, it says that rational thinkers are in a position to know something. The idea, already mentioned, is that rationality requires one to respect one's evidence, which one cannot expect to do without knowing what it is.

8.4 AN ARGUMENT FOR SAMENESS OF EVIDENCE

Let us analyse the argument for sameness of evidence in detail. For simplicity, we may concentrate on cases in which one is rational, possesses all the relevant concepts, and is currently reflecting on one's evidence and its implications; one is epistemically active enough to know whatever one is in a position to know about one's evidence. If the sceptic can show that under these conditions one's evidence is the same in the two cases, the anti-sceptic would have little to gain by insisting that it is different when one is less epistemically active.

We start with the premise that one knows what one's evidence is. 'Evidence' here and throughout means one's total body of evidence. To know what one's evidence is in the relevant sense, one must do better than merely to think of it as 'my evidence'. To be in a position to respect one's evidence, one must identify its specific content in a more perspicuous and intrinsic way. One need not compress the identification into a single item of knowledge. The content can be specified by a class of appropriate properties, each of which one knows one's evidence to have under some canonical specification of the property. We assume on behalf of the sceptic that for each appropriate property a unique canonical specification is given. Let us concede for the sake of argument that such a notion of the canonical can be worked out in detail. We may also assume that if a property is appropriate, so is its complement. The first premise is therefore:

(1) For any appropriate property π, in any case in which one's evidence has π, one knows that one's evidence has π.

If we wanted to generalize (1) and the rest of the argument beyond cases in which one is rational, possesses all the relevant concepts, and is currently reflecting on one's evidence and its implications, we could replace 'knows' by 'is in a position to know'. One might wonder whether (1) generates an infinite regress, as Richard Fumerton (2000) has suggested. It does if being known to have π counts as an appropriate property whenever π does, but defenders of (1) should not concede that assumption. The appropriate properties are intrinsic to the content of one's evidence; being known to have such a property need not itself be intrinsic to the content of the evidence.

Whereas the first premise concerns first-personal knowledge of one's own case, the second concerns third-personal knowledge of one case from within another. For the argument to work, in the bad case one must know what one's evidence would be if one were in the good case, where the good case is presented descriptively. We can quite fairly assume that the terms 'the good case' and 'the bad case' abbreviate descriptions in which, for each appropriate property, if one's evidence in a case has an appropriate property then that is specified in the description of the case; likewise if one's evidence lacks an appropriate property. For the sceptic will insist that however much information one has about what would be so if one were in a given case, that still does not enable one to work out which case one is in. We may assume that one can refer to the appropriate properties, for that is already implicit in (1): if one's evidence has the appropriate property π, then one knows that it has π and so can refer to π; if it lacks π, then it has the appropriate complementary property not-π, so one knows that it has not-π, so one can refer to not-π, so one can refer to π. Thus one can attain trivial conceptual knowledge in the bad case about the appropriate properties of one's evidence in the good case simply by unpacking one's descriptive concept of 'the good case':

(2) For any appropriate property π, if in the good case one's evidence lacks π, then in the bad case one knows that in the good case one's evidence lacks π.

The third premise articulates the badness of one's predicament in the bad case. From premises each of which one knows in the bad case, one cannot deduce that one is not in the good case.

(3) It is consistent with what one knows in the bad case that one is in the good case.

Now restrict 'π' to appropriate properties and assume:

(4) In the bad case one's evidence has π.

Suppose further, as an assumption for reductio ad absurdum:

(5) In the good case one's evidence lacks π.

Premises (2) and (5) entail:

(6) In the bad case one knows that in the good case one's evidence lacks π.

Premises (1) and (4) entail:

(7) In the bad case one knows that one's evidence has π.

From 'In the good case one's evidence lacks π' and 'One's evidence has π' one can deduce 'One is not in the good case'. By (6) and (7), in the bad case one knows each premise of that deduction; hence:

(8) It is inconsistent with what one knows in the bad case that one is in the good case.

Now (8), which rests on assumptions (1), (2), (4), and (5), contradicts (3). Thus on assumptions (1)–(4) we can deny (5) by reductio ad absurdum:

(9) In the good case one's evidence has π.

We can conditionalize (9) on assumption (4):

(10) If in the bad case one's evidence has π, then in the good case one's evidence has π.

Here (10) rests on assumptions (1)–(3). Since the appropriate properties were assumed to be closed under complementation, we can run through the argument (1)–(10) with 'not-π' in place of 'π', yielding:

(11) If in the bad case one's evidence has not-π, then in the good case one's evidence has not-π.

Contraposition on (11) yields the converse of (10). Therefore, generalizing on 'π' in (10) and (11), we have:

(12) One's evidence in the good case has the same appropriate properties as one's evidence in the bad case.

The conclusion (12) rests on assumptions (1), (2), and (3). It may be restated as the claim that one's evidence is the same in the good and bad cases, where evidence is individuated by the appropriate properties. If something like this argument is not the reason for which sceptics and others think that one has the same evidence in the two cases, it is not at all clear what is.

8.5 THE PHENOMENAL CONCEPTION OF EVIDENCE

That one has the same evidence in the good and bad cases is a severe constraint on the nature of evidence. It is inconsistent with the view that evidence consists of true propositions like those standardly offered as evidence for scientific theories. For example, the good case in which I see that the dial reads 0.407 corresponds to a bad case in which the dial does not read 0.407 but I hallucinate and it is consistent with everything I know that the dial reads 0.407. Since the proposition that the dial read 0.407 is false in the bad case, it is not evidence in the bad case. If my evidence is the same in the two cases, then that the dial read 0.407 is not evidence in the good case either. For similar reasons, (12) does not permit my evidence to include perceptual states individuated in part by relations to the environment. No matter how favourable my epistemic circumstances, I am counted as having only as much evidence as I have in the corresponding sceptical scenarios, no matter how distant and bizarre. Retinal stimulations and brain states fare no better as evidence, for in some sceptical scenarios they are unknowably different too. Thus (12) drives evidence towards the purely phenomenal.

We should not assume ourselves to grasp the concept of the phenomenal quite independently of (12). Instead, the phenomenal may be postulated as comprising those conditions, whatever they are, which rational subjects can know themselves to be in whenever they are in them. Such conditions may be supposed to comprise conditions on present memory experience as well as on present perceptual experience (Lewis 1996: 553). That such conditions exist is supposedly guaranteed by the argument that rationality requires one to respect one's evidence and cannot require one to respect something unless one is in a position to know what it is.[3]

[3] Fumerton 2000 suggests an alternative conception of the phenomenal as that which supervenes on relations of direct acquaintance. Presumably, to be directly acquainted with something is to be acquainted with it but not by being acquainted with something else. But then we lack a sound argument to show that I cannot be directly acquainted with something in the good case—such as my hand—with which I am not directly acquainted in the bad case. Some of Fumerton's remarks also presuppose the equivalence of the notion of the phenomenal as what one is always in a position to know about with

The argument for (12) is not vulnerable to a distinction between relevant and irrelevant alternatives to the good case, for it in no way assumes the relevance of the bad case to the good case. It does not use the sceptical claim that it is consistent with what one knows in the good case that one is in the bad case; it uses only the uncontested claim (3) that it is consistent with what one knows in the bad case that one is in the good case. Although that may be to assume the relevance in some sense of the good case to the bad case, that assumption is uncontroversial, since the good case is the sort of case one believes oneself to be in and appears to oneself to be in if one is in the bad case. Even if in the good case one properly ignores the bad case, the argument to (12) still shows (given its premises) that one's evidence in the good case cannot exceed one's evidence in the bad case.

Does a distinction between relevant and irrelevant alternatives make trouble for the sceptic's further claim that false belief in the bad case precludes knowledge in the good case? Perhaps falsely believing *p* with given evidence in a case β precludes knowing *p* with the same evidence in a case α only if β is a relevant alternative to α in some sense of 'relevant' in which the bad case is not a relevant alternative to the good case. Although that is not the present issue, it is difficult not to feel sympathy for the sceptic here. If one's evidence is insufficient for the truth of one's belief, in the sense that one could falsely believe *p* with the very same total evidence, then one seems to know *p* in at best a stretched and weakened sense of 'know'. We might contrast it with a more robust sense in which one knows the evidence itself, if evidence can be conceived propositionally. But all these questions presuppose that one's evidence is indeed the same in the good and bad cases. How compelling is the argument for (12)? In particular, how compelling is the justification of its crucial premise (1)?

8.6 SAMENESS OF EVIDENCE AND THE SORITES

We can undermine the argument for (12), and in particular its crucial premise (1), by constructing a parallel argument from (1) to a clearly false conclusion. Whatever the nature of evidence, rational thinkers do not always know what their evidence is. The argument exploits ordin-

a notion of the phenomenal as what one is infallible about. These notions are not obviously equivalent. I could be in a position to know *p* while falsely believing ~*p*, because my guru tells me ~*p*. If one can have contradictory beliefs, I might even know *p* while deceiving myself into believing ~*p*.

ary limits to one's powers of discrimination. It is an application of the anti-luminosity argument of section 4.3, with some modifications to clarify its relation to the argument presented in section 8.4. The argument shows that the condition that one's evidence has the appropriate property π is not luminous; it can obtain even when one is not in a position to know that it obtains.

Let t_0, t_1, t_2, . . ., t_n be a long sequence of times at one-millisecond intervals. Imagine that one's experience very gradually changes from t_0 to t_n; for example, one watches the sun slowly rise. One loses exact track of time. One's evidence at the beginning of the process (pitch darkness) is quite different from one's evidence at the end (bright daylight). Some of the appropriate properties of one's evidence are different; for purposes of this argument, it does not matter whether the appropriate properties exhaust the content of one's evidence. We may assume that the complement of an appropriate property is itself an appropriate property, although the purpose of the argument could be achieved without that assumption. For $0 \leq i \leq n$, let 'α_i' abbreviate a description of the case one is in at t_i; the description specifies the time t_i in clock terms and lists the appropriate properties which one's evidence then has and those which it then lacks. As with the sceptic's original argument, we may assume that one can refer to the appropriate properties, for that is implicit in (1). Thus one can attain trivial conceptual knowledge in one case about the appropriate properties of one's evidence in another case simply by unpacking one's descriptive concept of the latter case; in particular:

(2_i) For any appropriate property π, if in α_{i-1} one's evidence lacks π, then in α_i one knows that in α_{i-1} one's evidence lacks π.

The justification of (2_i) is just like the justification of (2) above.

Now consider the description of what is in fact the case one was in a millisecond ago. Given one's limited powers of discrimination, one does not know propositions from which one can deduce that that description does not apply to one's own case:

(3_i) It is consistent with what one knows in α_i that one is in α_{i-1}.

Since the purposes of this chapter require only one example in which (1) has false consequences, any readers lucky enough to have perfect discrimination amongst their own states should consider the less fortunate example of the present author, who is frequently in a predicament like (3_i). In such cases, (3_i) is obvious in roughly the way in which it is obvious that it is consistent with what I know by sight when I am in fact looking at a distant tree *i* millimetres high that I am looking at a tree only $i-1$ millimetres high. From premises which I know on the basis of sight to the conclusion that I am not looking at a tree only $i-1$ millimetres high, there is no hope of constructing a valid deduction, not even one which I am somehow not in a position to carry out. Similarly, from premises which I know in α_i to the conclusion that I am not in α_{i-1}, there is no hope of constructing a valid deduction, not even one which I am somehow not in a position to carry out.

The argument proceeds as before. Restrict 'π' to appropriate properties and assume:

(4_i) In α_i one's evidence has π.

Suppose further, as an assumption for reductio ad absurdum:

(5_i) In α_{i-1} one's evidence lacks π.

Premises (2_i) and (5_i) entail:

(6_i) In α_i one knows that in α_{i-1} one's evidence lacks π.

Premises (1) and (4_i) entail:

(7_i) In α_i one knows that one's evidence has π.

From 'In α_{i-1} one's evidence lacks π' and 'One's evidence has π' one can deduce 'One is not in α_{i-1}'. By (6_i) and (7_i), in α_i one knows each premise of that deduction; hence:

(8_i) It is inconsistent with what one knows in α_i that one is in α_{i-1}.

Now (8_i), which rests on assumptions (1), (2_i), (4_i), and (5_i), contradicts (3_i). Thus on assumptions (1) and (2_i)–(4_i) we can deny (5_i) by reductio ad absurdum:

(9_i) In α_{i-1} one's evidence has π.

We can conditionalize (9_i) on assumption (4_i):

(10_i) If in α_i one's evidence has π, then in α_{i-1} one's evidence has π.

Here (10_i) rests on assumptions (1), (2_i), and (3_i). Since the appropriate properties were assumed to be closed under complementation, we can run through the argument (1)–(10_i) with 'not-π' in place of 'π', yielding:

(11_i) If in α_i one's evidence has not-π, then in α_{i-1} one's evidence has not-π.

Contraposition on (11_i) yields the converse of (12_i). Thus, generalizing on 'π' in (10_i) and (11_i), we have:

(12_i) One's evidence in α_{i-1} has the same appropriate properties as one's evidence in α_i.

Proposition (12_i) rests on assumptions (1), (2_i), and (3_i). But the relation between the cases in (12_i) is transitive; if one's evidence in case β has the same appropriate properties as one's evidence in case γ and one's evidence in case γ has the same appropriate properties as one's evidence in case δ, then one's evidence in β has the same appropriate properties as one's evidence in δ, for what is in question is exact sameness in all properties from a fixed class. Although (3_i) claims only that α_{i-1} and α_i are indiscriminable, and indiscriminability is a non-transitive relation, we have deduced from it and the other premises the transitive relation of exact sameness of evidence in the appropriate respects. Thus (12_1), . . ., (12_n) together yield:

(13) One's evidence in α_0 has the same appropriate properties as one's evidence in α_n.

The conclusion (13) rests on assumptions (1), (2_1), . . ., (2_n), (3_1), . . ., (3_n). But (13) is obviously false. One's evidence at the end of the process is grossly different from one's evidence at the beginning; it differs in many of its appropriate properties. Since (2_1), . . ., (2_n), (3_1), . . ., (3_n) are true, for reasons already given, (1) is false.

Even if we drop the assumption that the complements of appropriate properties are themselves appropriate, we still have the argument to (10_i), and therefore by transitivity to the conclusion that if in α_n one's evidence has an appropriate property, then in α_0 one's evidence already had that property. That is obviously false, too. One does not always know the appropriate properties of one's evidence; one does not always know what one's evidence is.

To the objection that the argument is undermined by its obvious similarity to a sorites paradox, the reply is just as in section 4.5, and will not be repeated here. In brief, the argument in a sorites paradox has an obviously false premise when the vague terms at issue are sharpened; here that is not so.

Fumerton (2000) points out that the sorites argument would show that (1) can fail in a small way; it would not show that (1) can fail in a large way, as is held to occur in the bad case. However, one's evidence in the bad case can appear exactly similar to one's evidence in the good case, not because it is almost exactly similar, but because it is so radically impoverished that one lacks evidence of its impoverishment. Moreover, the usual reasons for claiming that one is always in a position to know exactly what one's evidence is do not naturally evolve into reasons for claiming that one is *always* in a position to know approximately what one's evidence is. We are *often* in a position to know approximately what our evidence is; that our position should occasionally be much worse than that, as in the bad case, is no surprise.[4]

8.7 THE NON-TRANSPARENCY OF RATIONALITY

The argument against (1) does not depend on any specific theory of evidence. The crucial assumption about evidence is just that its appropriate properties can vary between the endpoints of a spectrum of cases, as they must if we are to learn from experience. *Whatever* evidence is, one is not always in a position to know what one has of it. Thus nothing would be gained by a retreat to the fallback claim that one always knows (or is in a position to know) what one's evidence *appears* to be. For we can replace the words 'one's evidence has [lacks] π' in the preceding argument by 'one's evidence appears to have [lack] π'. Under this modification, (1) expresses the fallback claim, (2_1), . . ., (2_n) can be justified in the same way as before, (3_1), . . .,(3_n) are unchanged, and (13) remains hopelessly implausible, so the argument refutes the fallback claim, too. One does not always know what one's evidence appears to be.

If the phenomenal is postulated as comprising those conditions of the subject, whatever they are, which are accessible to the subject whenever they obtain, and therefore satisfy something like desideratum (1) for evidence, then the phenomenal is empty. We have the illusion of coming ever closer to a phenomenal core of experience by progressively eliminating every feature which can fail to be accessible to the subject, but, like the sequence of open intervals (0,1), (0,1/2), (0,1/4), . . ., this sequence of approximations converges to the empty set.

We could modify (1) by relativizing appropriateness to cases. The modified variant of (1) would claim that, for any case α and any property π appropriate to α, if in α one's evidence has π, then one knows in α that one's evidence has π. We could then no longer argue to (13), because 'appropriate' in the modified (12_i) would have different relativizations for different values of *i*. But the argument for sameness of evidence in the good and bad cases would fail, for although we could

[4] Following Poincaré, Russell used the non-transitivity of indistinguishability in sensation to argue for imperceptible differences amongst our sense data (Russell 1993: 148, originally published in 1914). A. J. Ayer replied that the only notion of exact resemblance applicable to sense data is equivalent to the relation of apparent exact resemblance between material things, which can be non-transitive (1940: 132–4). That reply provides no basis for resistance to the arguments of this chapter.

show that one's evidence in the good case had the same properties appropriate to the *bad* case as one's evidence in the bad case, we could not show that one's evidence in the good case had the same properties appropriate to the *good* case as one's evidence in the bad case. Indeed, we could not show that one was always in a position to know which properties of evidence were appropriate to one's own case. The proposed relativization plays into the hands of the present strategy.

The problem remains: how can rational thinkers respect their evidence if they do not know what it is? If rationality requires one to respect one's evidence, then it is irrational not to respect one's evidence. But how can failing to respect one's evidence be irrational when one is not in a position to know that one is failing to respect one's evidence? More generally, how can ϕ-ing be irrational when one is not in a position to know that one is ϕ-ing?

The standard conception of rationality depends on a distinction between the *aims* and *methods* of cognitive activity. On that conception, truth is an aim. We cannot attain it directly; we cannot follow the rule 'Believe truly!' when we do not know what is true. Therefore we must use methods to reach the truth. Rationality is a method. We can follow rules of rationality because we are always in a position to know what they require. If the argument of section 8.6 is correct, this picture of rationality is mistaken. Just as one cannot always know what one's evidence is, so one cannot always know what rationality requires of one. Just like evidence, the requirements of rationality can differ between indiscriminable situations. Rationality may be a matter of doing the best one can with what one has, but one cannot always know what one has, or whether one has done the best one can with it. If something is a method only if one is always in a position to know whether one is complying with it, then there are no methods for learning from experience. But that standard is too exacting to be useful. We can use something as a method in contexts in which one is usually in a position to know whether one is complying with it, even if in other contexts one is not usually in a position to know whether one is complying with it. In that sense, we can use even believing truly as a method in contexts in which one is usually in a position to know what is true: for example, when forming beliefs in normal conditions about the spatial arrangement of medium-sized objects in one's immediate environment. In more difficult contexts, believing truly becomes an aim and we fall back on the method of believing rationally. Rationality becomes a sub-goal on the way to truth. That does not require one always to be in a position to know what rationality requires of one; it requires merely that one often knows what rationality requires when one does not know what truth requires. Nothing has been said here to undermine that requirement. In still more problematic contexts, paradoxes throw our very standards of rationality into doubt, and we fall back still further on what workable methods we can find. Cognition is irremediably opportunistic.

There is a pragmatist and subjective Bayesian project to operationalize epistemology by working only with concepts whose application is always accessible to the agent. The argument of this chapter implies that the project is doomed to failure.

Uncertainty about evidence does not generate an infinite regress of evidence about evidence about . . .. In order to reflect adequately on one's evidence, one might need evidence about one's evidence, and in order to reflect adequately about the latter evidence, one might need evidence about it, and so on. But this regress is merely and harmlessly potential. We cannot in fact realize infinitely many levels of adequate reflection; at best, further reflection enables us to realize finitely many further stages. At some stage one must rely on unreflective causal sensitivity to evidence (see section 9.3).

One can be causally sensitive to a factor without being in a position to have exact knowledge of it, as when one is causally sensitive through unaided perception to the distances between objects in one's environment. One can be causally sensitive to appropriate properties of one's evidence without being in a position to know them exactly. Causal sensitivity need not be perfect to be genuine. Sufficiently bad cognitive circumstances may involve obstacles even to causal sensitivity to one's evidence. The bad case in a sceptical argument may be a case in point. One's cognitive circumstances may be so bad that one is in no position to know how impoverished one's evidence is in comparison to the good case. Our causal insensitivity to any difference in evidence between the two cases does not show that there is no difference in evidence between them.

It has not been shown that the good and bad cases do differ in evidence. That requires a positive account of evidence, which Chapters 9 and 10 will develop. They defend the view that one's total evidence (not one's evidence for *p* alone) is simply one's total knowledge, on which the assumption that one has the same total evidence in the two cases is tantamount to the sceptic's conclusion. For since, uncontroversially, in the bad case one fails to know *p*, *p* would not be part of one's total evidence in the bad case, and would therefore not be part of one's total evidence in the good case either; so in the good case, too, one would not know *p*. A sceptic who assumes that one's total evidence is the same begs the question against a non-sceptic who takes that view of evidence. Of course, the argument of this chapter does not assume the equation of one's total evidence with one's total knowledge; rather, it lays the groundwork for the equation.[5]

For present purposes, what matters is that sameness of evidence has not been established, and a salient argument for it has turned out to rest on a false premise. For all that the sceptic has shown, one has more evidence in the good case than in the bad case, and knowledge in the former is unthreatened by false belief in the latter.

The problem is not confined to the sceptic. It also affects those non-sceptics who argue that in the good case we know *p* even though we falsely believe *p* in some cases in which we have the same evidence as in the good cases, because those bad cases are irrelevant (at least in this context). Such theorists have not eliminated the hypothesis that one knows *p* only if one does not falsely believe *p* on the same evidence in any case at all, relevant or irrelevant. That hypothesis does not entail scepticism.

Contextualists may argue that the extension of 'evidence' waxes and wanes with the context of utterance just as they suppose the extension of 'knowledge' to do. But then they cannot use 'the same evidence' as a fixed standard against which to measure contextual variation in standards of relevance. 'One knows *p*' is supposed to count as true in the good case when the bad case is irrelevant; but if (speaking in such a context) the bad case also counts as differing from the good case in non-pragmatic respects such as evidence, why invoke pragmatic respects such as relevance?

8.8 SCEPTICISM WITHOUT SAMENESS OF EVIDENCE

Sometimes the good and bad cases in a sceptical argument have a different structure from that considered so far. Scepticism about *p* does not always require the (metaphysical) possibility of a bad case in which one falsely believes *p*. Let *p* be a mathematical truth, and therefore a necessary truth. Thus no case in which one falsely believes *p* is possible; yet one can still doubt *p*, by doubting the reliability of the methods which led one to believe *p*. After all, someone with great faith in a certain coin might decide to believe *p* if it comes up heads and to believe ~*p* if it

[5] The equation allows sameness of evidence between two cases in which, without knowing *p*, one justifiably believes *p*, in one case truly, in the other falsely. Fumerton 2000 describes such a case, under the impression that it constitutes a difficulty for the equation.

comes up tails; if he believes *p* because the coin came up heads, he does not know *p*, although he could not have believed *p* falsely. His belief fails to be knowledge because the method by which he reached it could just as easily have led to a false belief in a different proposition (see also section 4.4). His evidence in the bad case includes the proposition that the coin came up tails, and therefore differs from his evidence in the good case. Even when false belief in *p* is possible, one's evidence in the bad case which motivates scepticism about *p* need not be the same as in the good case. Incoherent dreams feel coherent; one might therefore doubt the coherence of one's present experience, even though it feels, and in fact is, coherent. Since one's experience is coherent in the good case and incoherent in the bad case, one's evidence presumably differs between the two cases. The sceptic need not claim that in some possible case, one's experience is incoherent and one has the same evidence which one actually has, for that would be too close to asserting dogmatically that one's actual experience is incoherent. Rather, the locus of the doubt is the method by which one reaches the belief that one's experience is coherent.

Does the discussion of sceptical arguments in sections 8.2–8.7 generalize to these examples? One might think not. 'Method' has replaced 'evidence' as the crucial term, and they seem to be crucially disanalogous: we are far more strongly tempted to assume that one is always in a position to know what one's evidence is than to assume that one is always in a position to know what method one is using. The sceptic will happily allow that our beliefs may have inaccessible, unconscious causes, and argue that for all we know such causes are quite insensitive to whether the beliefs they cause are true. What if we are in fact using a method which cannot yield false beliefs? The sceptic will point to cases in which we merely appear to be using that method and our resulting beliefs are false. Such cases are supposed to falsify our knowledge claims. For the sceptic, the methods on whose reliability the epistemic status of our beliefs depends are individuated by appearances; the falsity of beliefs reached in cases which appear the same as the actual case in respect of one's method constitutes unreliability in one's actual apparent method.

In effect, the sceptic distinguishes the *process* by which one's belief was caused from the *rule* which one used in reaching the belief. Processes are at the subpersonal level, rules at the personal level. One can take responsibility for one's rules in a way in which one cannot for the processes. One's rationality depends on the rules which one uses rather than the processes which go on in one. One is typically not in a position to know what process caused one's belief, but we are tempted to suppose that one is always in a position to know what rules one used in reaching it. For if one was not in a position to know what rules one was using, and one's rationality depended on the rationality of one's rules, how could one be required to be rational? The requirement that one is always in a position to know what rules one is using forces us into individuating them phenomenally, just as the corresponding epistemic requirement on evidence forced us into individuating it phenomenally. If a rule were individuated non-phenomenally, one could to all appearances be using it while not in fact be doing so; in which case one would not be in a position to know that one was not using that rule. An argument similar to (1)–(12) concludes that one is using the same rule in the good and bad cases. According to the sceptic, what matters for knowledge is the reliability of the rule, not of the process, so one's false belief in the bad case makes one's rule in the good case unreliable, and therefore undermines knowledge in the good case.

The sceptic's conception of a rule collapses. By an argument parallel to (1)–(13) (via (2_i)–(12_i)), only trivial rules meet the epistemic requirement. For a series of indiscriminable differences links a case in which one uses a given rule to a case in which one uses a quite different rule. For example, one initially believes *p* for reason R while giving no weight to reason R*; gradually one gives less weight to R and more to R*, until finally one believes *p* for reason R* while giving no weight to R. R and R* differ so much in kind that believing for reason R and believing for reason R* amount to using different rules. An argument just like that of section 8.6 refutes the assumption that in every case one is in a position to know what rule one is using. Even when the sceptic does not assume identity of evidence between the good and bad cases, the underlying dialectic is the same.

We can fall into scepticism if we attribute too much self-knowledge to the subject in bad cognitive circumstances, for the asymmetry in knowledge between the good and bad cases requires an asymmetry in self-knowledge. Once we relax our claims to self-knowledge, we strengthen our claim to knowledge of the external world. Sceptical arguments fail when they depend on exempting an internal world of appearances, for they depend on misconceiving appearances as just what they appear to be. The ruthless sceptic grants no exemptions. If the sceptic must argue that we never even know how things appear to us, should we still harbour the sneaking suspicion that scepticism is right after all?

9
Evidence

9.1 KNOWLEDGE AS JUSTIFYING BELIEF

Tradition has it that the main problems of philosophy include the nature of knowledge. But, in recent decades, questions of knowledge seem to have been marginalized by questions of justification. Thus, according to Crispin Wright,

> knowledge is not really the proper central concern of epistemologico-sceptical enquiry. . . . We can live with the concession that we do not, strictly, *know* some of the things we believed ourselves to know, provided we can retain the thought that we are fully justified in accepting them. (1991: 88; Wright's italics)

Similarly, John Earman argues that accounts of knowledge are irrelevant to the philosophy of science, because in it 'the main concern is rarely whether or not a scientist 'knows' that some theory is true but rather whether or not she is justified in believing it' (1993: 37).[1] Once Gettier showed in 1963 that justified true belief is insufficient for knowledge, and therefore that knowledge is unnecessary for justified true belief, it became natural to ask: if you can have justified true beliefs, why bother with knowledge?[2]

The argument of Chapter 3 indicated that if one is disposed to respond rationally to future evidence, then one's future prospects are better if one now has knowledge than if one now has mere justified true belief. But even if we restrict the comparison between knowledge and justified true belief to what they do for one in the present, there is still a lacuna in the case for the unimportance of knowledge. Grant, for the sake of argument, that knowledge is important now only if it is some-

[1] Earman is discussing externalist accounts of knowledge, but the quoted comment would clearly apply to internalist accounts too. Earman's further point, that 'because science is a community enterprise the only forms of justification that are scientifically relevant are those which are stateable and open to public scrutiny', may be most relevant to externalist accounts. See also Craig 1990b: 272. For the contrary view that scepticism about knowledge entails scepticism about rationality and justification, see Unger 1975: 197–249.

[2] Kaplan 1985 argues along similar lines.

how essential to the present justification of belief.[3] Although it has been shown that *what is justified* need not be knowledge, even when it is true, it has not been shown that *what justifies* need not be knowledge. Only one end of the justification relation has been separated from knowledge. Suppose that knowledge, and only knowledge, justifies belief. That is, in any possible situation in which one believes a proposition p, that belief is justified, if at all, by propositions $q_1, \ldots, q_n$ (usually other than p) which one knows. On that supposition, if justified belief is central to epistemologico-sceptical inquiry and the philosophy of science, then so too is knowledge. Now assume further that what justifies belief is *evidence* (this assumption is briefly discussed in section 9.8). Then the supposition just made is equivalent to the principle that knowledge, and only knowledge, constitutes evidence. This chapter defends that principle; it equates S's evidence with S's knowledge, for every individual or community S in any possible situation. Call this equation $E = K$.[4]

As usual, 'knowledge' is understood as propositional knowledge. The communal case is needed: science depends on public evidence, which is neither the union nor the intersection of the evidence of each scientist. We can ascribe such knowledge by saying that p is known in community S, or that we know p, which is not equivalent to saying that some, many, most, or all of us know p.

The proposed account uses the concept of knowledge in partial elucidation of the concepts of evidence and justification. To some people it will therefore seem to get things back to front. For although knowledge is more than justified true belief, many philosophers still expect to use concepts such as *evidence* and *justification* in a more complex explanation of the concept *knows*; it would then be circular to use the latter to explain the former. Others prefer to use concepts of a different kind, such as *causation* or *reliability*, to explain the concept *knows*; but even they are likely to regard the concept *knows* as so much in need of explanation itself that its pre-theoretic use would lack explanatory value.

That order of explanation has been reversed in this book. The concept *knows* is fundamental, the primary implement of epistemological inquiry. Chapter 1 rejected the programme of understanding knowledge in terms of the justification of belief. That frees us to try the experiment

[3] Wright speaks of justified acceptance rather than belief; although Earman speaks of belief, some philosophers of science contrast it with acceptance and regard the latter as a more appropriate attitude towards scientific theories. The distinction is not important here.

[4] The principle is stated, and applied to the problem of vagueness, in Williamson 1994b: 245–7.

of understanding the justification of belief in terms of knowledge. Of course the concept *knows* is vague; so is the concept *justified*.

For those who remain sympathetic to the orthodox order of explanation, some more irenic points can be made. The equation E = K could be true without being knowable a priori. As a universal generalization over all metaphysically possible situations, it is necessarily true, if true at all (by the S4 principle that a necessary truth is necessarily necessary); but we cannot presume that a necessary truth is knowable a priori. E = K equates the extensions of the concepts *knowledge* and *evidence* in any possible situation; that is enough to make it an informative thesis. By itself, E = K does not equate the concepts themselves; nor is it to be read as offering an analysis of either the concept *evidence* or the concept *knowledge*, or as making one concept prior to the other in any sense. Of course, in offering *arguments* of a broadly a priori kind for E = K, like those below, one commits oneself at least to its a priori plausibility; in the best case for those arguments, they would provide a priori knowledge of E = K. But even if the concepts are equivalent a priori, it does not follow that one is prior to the other.

More positively, we may speculate that standard accounts of justification have failed to deal convincingly with the traditional problem of the regress of justifications—what justifies the justifiers?—*because* they have forbidden themselves to use the concept *knowledge*. E = K suggests a very modest kind of foundationalism, on which all one's knowledge serves as the foundation for all one's justified beliefs. Perhaps we can understand how something could found belief only by thinking of it as knowledge.

9.2 BODIES OF EVIDENCE

When is e evidence for the hypothesis h, for a subject S? Two conditions seem to be required. First, e should speak in favour of h. Second, e should have some kind of creditable standing. At least as a first approximation, we can model the first condition in probabilistic terms: e should raise the probability of h. That is, the probability of h conditional on e should be higher than the unconditional probability of h; in symbols, $P(h|e) > P(h)$. The conditional probability $P(h|e)$ is defined as the ratio $P(h \wedge e)/P(e)$ when $P(e) \neq 0$, and is otherwise undefined. Thus the condition that $P(h|e) > P(h)$ obtains if and only $P(h \wedge e) > P(h)P(e)$. What kind of probability is P? It is not a priori, for whether e raises the probability of h may depend on background information. For example, the proposition that John belongs to a certain club might raise the probability that he is single relative to the background information that it is a club for singles, but lower it relative to the background information that it is a club for spouses. However, *e* itself should not be built into the background information, for that would give P(*e*) the value 1, in which case P(*h*|*e*) and P(*h*) would be equal and *e* would not be evidence for anything. Let us leave the nature of P underspecified until the next chapter. Now, *e* may raise the probability of *h* in the sense that P(*h*|*e*) > P(*h*) even if S knows that *e* is false or has no idea whether *e* is true; but then, for S, *e* would not be evidence for *h*. That is why we need the second condition, that *e* should have a creditable standing. A natural idea is that S has a *body of evidence*, for use in the assessment of hypotheses; that evidence should include *e*. The probability distribution P is informed by some but not all of S's evidence. We can therefore formulate a simple schematic proposal:

EV *e* is evidence for *h* for S if and only if S's evidence includes *e* and P(*h*|*e*) > P(*h*).

One consequence of EV is that *e* is evidence for *h* only if *e* is evidence for itself. For if P(*h*|*e*) > P(*h*), then P(*e*) is neither 0 (otherwise P(*h*|*e*) is ill defined) nor 1 (otherwise P(*h*|*e*) = P(*h*)). Hence P(*e*|*e*) is well defined with the value 1, which is greater than P(*e*), so *e* is evidence for *e*, by EV with '*e*' substituted for '*h*'. But is it not circular for anything to be evidence for itself? A critic might therefore argue that one's evidence does not consist of a fixed body of propositions; either it depends on the hypothesis under assessment, where no proposition belongs to the evidence relative to its own assessment, or it does not consist of propositions.

The critic is not entitled to assume without argument that classifying *e* as evidence for itself involves circularity of any vicious kind. Certainly EV does not make it trivially easy to have evidence for *e*, for *e* is evidence for itself for S only if S's evidence includes *e*. By E = K, that requires S to know *e*, which may not be easy. The result that *e* is evidence for itself may be as harmless as the consequence of a standard definition of provability in a formal system that every axiom has a one-line proof, consisting of the axiom itself. Of course, if someone asks 'What is the evidence for *h*?', one is not expected to cite *h* itself, but the reason might be that it would be conversationally inappropriate rather than false to do so. In answer to the question 'Who lives in the same house as Mary?' it would be conversationally inappropriate to cite Mary herself; nevertheless, it is true that Mary lives in the same house as Mary (Grice 1989). The question 'What is the evidence for *h*?' is often a challenge to the epistemic standing of *h* and related propositions. In some contexts the challenge is local, restricted to propositions derived in some way from *h*. In other contexts the challenge is global, extending to all propositions with the same kind of pedigree as *h*. In answering the question, one is expected not to cite propositions under challenge, since their status as evidence has been challenged. Thus when the question 'What is the evidence for *e*?' is meant as a challenge to the epistemic standing of *e*, one is expected not to cite *e* in response.

Could we treat the claim that *e* is evidence for itself as false rather than conversationally inappropriate by treating 'evidence' as a context-sensitive term? The idea would be that the question 'What is the evidence for *e*?', meant as a challenge, creates a context in which *e* falls outside the extension of 'S's evidence'. But that seems too drastic. For example, suppose that a doctor asks you, 'Do you feel a tingling sensation?' and you answer, 'No.' If you were asked 'What is your evidence for the proposition that you do not feel a tingling sensation?', you might be at a loss to answer, for the question seems to expect some *further* evidence for the proposition, and you might look in vain for such further evidence. Nevertheless, when we assess the status of your claim that you did not feel a tingling sensation on your evidence, we do not exclude that proposition from your evidence. Its presence justified your claim. This is not to deny that the extension of 'evidence' may vary slightly with context, perhaps corresponding to slight contextual variation in the extension of 'knowledge' (and therefore, presumably, in the extension of 'mental state' too, by section 1.4). The point is just that challenging *e* by itself is not enough to exclude *e* from the extension of 'evidence'.

One sceptical strategy is to exploit the dialectical effects of challenging propositions. If one is never entitled to rely on something under challenge, one will very soon be left with very little. For example, the sceptic can challenge the belief that there are good reasons, and then charge any attempt to provide a good reason for it with begging the question. We should be sceptical of such a sceptic's reliance on the power of challenge. The sceptic relies uncritically on rules of dialectical engagement which evolved to serve more practical purposes, without questioning their appropriateness to the radical questions which scepticism raises. If challenging something thereby makes it dialectically unusable, then the power of challenge might hinder rather than help the pursuit of truth if it is not used with restraint. By refusing to associate questions of evidence too closely with questions of dialectical propriety, we can preserve EV.

EV concerns the evidence-for relation, as do most discussions of evidence.[5] The focus of this chapter is elsewhere. It concerns the nature of the first relatum *e* of the evidence-for relation rather than its relation to the second relatum *h*. Whether EV needs revision will be left open; the present aim is to investigate its constituent 'S's evidence includes *e*'. Chapter 10 develops a theory of evidential probability to address the relation between evidence and what it supports.

Why does it matter what counts as evidence? Consider the idea that one should proportion one's belief in a proposition to one's evidence for it. How much evidence one has for the proposition depends on what one's evidence is. More precisely, a theory of evidence is needed to give bite to what Carnap calls *the requirement of total evidence*:

> [I]n the application of inductive logic to a given knowledge situation, the total evidence available must be taken as a basis for determining the degree of confirmation. (1950: 211; compare Hempel 1965: 63–7)

If too much or too little is counted as evidence, inductive principles will be misapplied. Given the requirement of total evidence, disputes between different theories of evidence are not merely verbal; they involve disagreements as to which inductive conclusions are warranted. Formulations of the total evidence requirement in terms of knowledge encourage E = K, which identifies the total evidence available with the total knowledge available. For example, Peirce writes:

> I cannot make a valid probable inference without taking into account whatever knowledge I have (or, at least, whatever occurs to my mind) that bears on the question. (1932: 461)

Carnap himself describes the evidence as (observational) knowledge. Given E = K, the original idea becomes something like this: one should proportion one's belief in a proposition to the support which it receives from one's knowledge.

The total evidence available must not be built into the probability distribution P in EV, otherwise no part of that evidence could confirm any hypothesis. In general, the total evidence must not be taken as a basis for determining the degree to which an individual piece *e* of that evidence increases the confirmation of a hypothesis *h*, for if *e* is part of the total evidence available then the confirmation of *h* prior to the acquisition of the total evidence presently available is also relevant. This is a form of the problem of old evidence (Glymour 1980: 85–93, Earman 1992: 119–35, Howson and Urbach 1993: 403–8, Maher 1996), which is

[5] The papers in the representative collection Achinstein 1983, for example, are largely concerned with questions about the evidential relation at the expense of questions about its first relatum.

discussed in the next chapter. But that does not undermine the point that the total evidence now available must be taken as the basis for determining the degree to which *h* is now confirmed in the non-comparative sense.

Theories of evidence also play a role when theses of the underdetermination of theory by data are assessed, for the data in question are the actual or potential evidence. If too much or too little is counted as evidence, then the standard for underdetermination is set uninterestingly high or uninterestingly low. E = K implies that undetermination theses of the relevant kind must count all knowable facts as data. Although this condition does not automatically make any argument for underdetermination circular, it is not easily met. Consider, for example, the underdetermination thesis that the theoretical facts do not supervene on the evidential facts: two possible worlds can differ in the former without differing in the latter.[6] One cannot establish this claim just by showing that the theoretical facts do not supervene on the facts which are in some sense observable; one must also show that they do not supervene on all the knowable facts. The gap would be filled by an argument that the knowable facts supervene on the observable facts, for then whatever failed to supervene on the observable facts would fail to supervene on the knowable facts too, by the transitivity of supervenience. But any such argument risks begging the question against the view that at least some theoretical facts are knowable.

9.3 ACCESS TO EVIDENCE

Chapter 8 argued that we are not always in a position to know what our evidence is. Consequently, a theory of evidence cannot be expected to provide a decision procedure which will always enable us to determine in practice whether our evidence includes a given item. In general, a philosophical theory of a concept is not required to provide a decision procedure which will always enable us to determine in practice whether it applies to a given item. The concept of evidence might have been expected to be special in this respect, for if it were problematic whether one's evidence included something, one would need evidence to decide whether one's evidence included it, and an infinite regress looms. It is therefore tempting to suppose both that it must be unproblematic

[6] Compare EI_3 in the useful classification of kinds of empirical indistinguishability in Earman 1993: 21.

whether one's evidence includes any given item, and that an adequate theory of evidence must explain how it manages to be so unproblematic. By the argument of Chapter 8, however, no correct theory of evidence can have that upshot. Certainly the equation E = K does not, but since that does not distinguish it from other theories of evidence, it constitutes no objection to E = K. In obvious symbolism, E = K equates E*p* and K*p*. The transparency of evidence would make E*p* equivalent to KE*p*. Given E = K, that is tantamount to making K*p* equivalent to KK*p*. But we saw in section 5.1 that K*p* does not entail KK*p*. This section explores our limited access to our evidence in the light of the equation E = K.

There is no infallible recipe for deciding in practice whether we know a proposition *p*. Sometimes we reasonably believe ourselves to know *p*, when in fact we do not know *p*, because *p* is false. Reputable authorities assert that Henry V died in 1422; I have no grounds for doubting them, and reasonably believe myself to know by testimony that Henry V died in 1422. But it is not inconceivable that he died in 1423, some elaborate conspiracy being responsible for present evidence to the contrary. According to E = K, if I do know that Henry V died in 1422, then my total evidence includes the proposition that Henry V died in 1422; but if Henry V died in 1423, then my belief that my total evidence includes that proposition is mistaken—my total evidence includes only the proposition that reputable authorities assert that Henry V died in 1422. E = K is an externalist theory of evidence, in at least the sense that it implies that one's evidence does not supervene on one's internal physical states. But if knowing is a mental state, as argued in Chapter 1, then one's evidence does supervene on one's *mental* states.

How does E = K avoid the threatened regress of evidence? The regress comes if evidence-based belief in a proposition *p* must always be preceded by evidence-based belief in a proposition about the evidence for *p*. We can distinguish two senses of 'evidence-based'. Call one's belief in *p* *explicitly* evidence-based if it is influenced by prior beliefs about the evidence for *p*. Explicitly evidence-based beliefs may be more common in science than in everyday life. Call one's belief in *p* *implicitly* evidence-based if it is appropriately causally sensitive to the evidence for *p*. A belief can be both explicitly and implicitly evidence-based. Now, explicitly evidence-based belief in *p* is not always preceded by *explicitly* evidence-based belief in a proposition about the evidence for *p*; this is consistent with E = K and most other theories of evidence. An explicitly evidence-based belief is influenced by a prior state of belief in a proposition about the evidence for *p*, and something has gone wrong if the latter belief is not at least implicitly evidence-based; but it need not be explicitly evidence-based. Thus there is no regress of explicitly evidence-based belief. There would be a different regress if implicitly evidence-based belief in *p* were always preceded by implicitly evidence-based belief in a proposition about the evidence for *p*. But the causal sensitivity of the belief in *p* to the evidence for *p* need not be mediated by further *beliefs* about the evidence for *p*. There need be no such beliefs.

How can a belief in *p* be implicitly evidence-based, if we are liable to misidentify the evidence for *p*? If the real evidence differs from the apparent evidence, will not the belief be causally sensitive to the latter rather than the former? But, as noted in section 8.7, we are liable to misidentify the apparent evidence, too. Causal sensitivity need not be perfect to be genuine. There can be a non-accidental rough proportionality between the strength of the belief and the strength of the evidence, even if distortions sometimes occur.

Similar questions arise about explicitly evidence-based belief. How can one follow the rule 'Proportion your belief in *p* to your evidence for *p*' when one doesn't know exactly what one's evidence is? Given E = K, the rule becomes 'Proportion your belief in *p* to the support that *p* receives from your knowledge': but is one not at best following the rule 'Proportion your belief in *p* to the support that *p* receives from *what you believe to be* your knowledge'? Consider an analogy. We can follow the rule 'Proportion your voice to the size of the room'. This is not because we are infallible about the size of the room. We sometimes make mistakes; but it does not follow that we are really following the rule 'Proportion your voice to *what you believe to be* the size of the room'. After all, it is often quite hard to know what beliefs one has about the size of a room; we are fallible in our beliefs about such beliefs. That one believes *p* is not a luminous condition. In general, if the fallibility of our beliefs about X posed a problem, it would not be solved by the move to our beliefs about our beliefs about X, because they are fallible too. But fallibility does not pose a problem here. To make a mistake in following a rule is not to follow a different rule. The rule is a standard of correctness for action, not a description of action. To have applied the rule 'Proportion your voice to the size of the room', one needs beliefs about the size of the room, but they need not have been true—although if they were false, one's application was faulty. Similarly, to have applied the rule 'Proportion your belief in *p* to the support that *p* receives from your knowledge', one must have had beliefs about how much support *p* received from one's knowledge, and therefore about one's knowledge, but those beliefs need not have been true—although if they were false, one's application was faulty.

None of this would be much consolation if our beliefs about our knowledge were hopelessly unreliable. Sceptics say that those beliefs have no rational basis, but they say the same about most of our other beliefs, too. We have found their reasons for saying so to be inadequate. Although we have no infallible procedure for determining whether we know *p*, in practice we are often in a position to know whether we know *p*.

The ways in which we decide whether we know *p* are not simply the ways in which we decide whether we believe that we know *p*. If I want to check whether I now really know that Henry V died in 1422, it would be relevant to return to my sources; it would be irrelevant to do that if I merely wanted to check whether I now really believe that I know that he died in 1422.

As Chapter 8 noted, alternative theories of evidence distort the concept in the attempt to make evidence something that we can infallibly identify. Characteristically, they interiorize evidence: it becomes one's present experience, one's present degrees of belief, or the like. Those attempts are quaint relics of Cartesian epistemology. Knowledge of the present contents of one's own mind is neither unproblematic nor prior to knowledge of other things. It is not obvious to me how many shades of blue I am presently experiencing, or to what degree I believe that there was once life on Mars. If one's evidence were restricted to the contents of one's own mind, it could not play the role that it actually does in science. The evidence for the proposition that the sun is larger than the earth is not just my present experiences or degrees of belief. If the evidence is widened to include other people's experiences or degrees of belief, or my past ones, then my identification of it becomes even more obviously fallible. In any case, that does not seem to be the right widening; it is more plausible that the evidence for a scientific theory is the sort of thing which is made public in scientific journals. If evidence is like that, our identification of it is obviously fallible.

9.4 AN ARGUMENT

Here is a schematic argument for E = K:

> All evidence is propositional.
> All propositional evidence is knowledge.
> All knowledge is evidence.
> ______________________________
> All and only knowledge is evidence.

The argument is obviously valid, but its premises are contentious. Its aim is simply to divide the contentiousness of the conclusion into manageable portions; sections 9.5, 9.6, and 9.7 respectively defend the three premises. Since 'knowledge' here means propositional knowledge, each premise follows from the conclusion; thus the conclusion is equivalent to the conjunction of the premises.

One's evidence is propositional if and only if it is a set of propositions. Propositions are the objects of propositional attitudes, such as knowledge and belief; they can be true or false; they can be expressed relative to contexts by 'that' clauses. For present purposes, we do not need a developed theory of propositions. If evidence is propositional, we can refer to evidence by using 'that' clauses: my evidence for the conclusion that the house was empty is *that* it was silent, *that* no lights were on in the evening, *that* the telephone went unanswered, ...

9.5 EVIDENCE AS PROPOSITIONAL

Why should all evidence be propositional? It would not be on a broad interpretation of 'evidence'. In the courts, a bloodied knife is evidence. It is natural to say that my evidence that I am getting a cold includes various sensations. Some philosophers apply the term 'evidence' to non-propositional perceptual states; Quine restricts it to the stimulation of sensory receptors (1969: 75). How can 'All evidence is propositional' do more than stipulate a technical use for the word 'evidence'?

Indiscriminate description of the ordinary use of a term and arbitrary stipulation of a new use are not the only options. We can single out theoretical functions central to the ordinary concept *evidence*, and ask what serves them. That strategy is pursued here. The argument below substantiates the familiar claim that only propositions can be reasons for belief (for example, Unger 1975: 204–6 and Davidson 1986; for opposing views, Moser 1989: 47–125 and Millar 1991). It also suggests a further conclusion: one grasps the propositions that are one's evidence; one can think them.

Consider inference to the best explanation (Harman 1965, Lipton 1991). We often choose between hypotheses by asking which of them best explains our evidence—which of them, if true, would explain the evidence better than any other one would, if true. Fossil evidence enables us to answer questions about terrestrial life in this way. Even if inference to the best explanation is not legitimate in all theoretical contexts, what matters for present purposes is that, where evidence does enable us to answer a question, a central way for it to do so is by inference to its best explanation. Thus evidence is the kind of thing which hypotheses explain. But the kind of thing which hypotheses explain is propositional. Therefore evidence is propositional.

The kind of thing which hypotheses explain is propositional. Inference to the best explanation concerns why-explanations, which can be put in the form '—— because . . .', which is ungrammatical unless declarative sentences, complements for 'that', fill both blanks. We cannot simply *explain Albania*, for 'Albania because . . .' is ill-formed. We can sometimes make sense of the injunction 'Explain Albania!', but only when the context allows us to interpret it as an injunction to explain why Albania exists, or has some distinctive feature. What follows 'why' is a declarative sentence, expressing the proposition to be explained—*that* Albania exists, or *that* it has the distinctive feature. It makes no significant difference if what is to be explained is one thing as contrasted with another (Lipton 1991: 75–98). For example, we may seek to explain why Kosovo rather than Bosnia was peaceful in 1995. The evidence in question would be the propositions that Kosovo was peaceful in 1995 and that Bosnia was not. The same goes for events: 'Explain World War I!' enjoins one to explain why it occurred, or had some distinctive feature. Again, the sensation in my throat is evidence for the conclusion that I am getting a cold in the sense that the hypothesis that I am getting a cold would best explain why I have that sensation in my throat. The evidence to be explained is *that* I have that sensation in my throat—not just that I have *a* sensation in my throat. Even in the courts, the bloodied knife provides evidence because the prosecution and defence offer competing hypotheses as to why it was bloodied or how it came into the accused's possession; the evidential proposition is *that* it was bloodied or *that* it came into the accused's possession. The knife is a source of indefinitely many such propositions.

One can use an hypothesis to explain why A only if one grasps the proposition that A. Thus only propositions which one grasps can function as evidence in one's inferences to the best explanation. By this standard, only propositions which one grasps count as part of one's evidence.

Similar points apply to explicitly probabilistic reasoning. If such reasoning can be assimilated to inference to the best explanation, or vice versa, so much the better. The best way of comparing the conditional probabilities of two hypotheses h and h^* on evidence e, $P(h|e)$ and $P(h^*|e)$, is often by calculating the inverse probabilities of e on h and h^*, $P(e|h)$, and $P(e|h^*)$. For example, a bag contains ten red or black balls; we wish to estimate how many of them are red; we are allowed to gain evidence only by sampling one ball at a time, noting its colour and replacing it. A good way to compare the probabilities of hypotheses about the number of red balls is by calculating the probabilities of the actual outcome *e* of the sampling (say, red fifteen times and black five times) on those hypotheses. One way of using those probabilities is to regard *h* as more probable than *h** given *e* ($P(h|e) > P(h^*|e)$) if and only if *h* makes *e* more probable than *h** does ($P(e|h) > P(e|h^*)$). Bayesians take this method to involve assigning the same prior probability to *h* and *h** ($P(h) = P(h^*)$); they treat as equally legitimate assignments of unequal prior probabilities to the hypotheses—perhaps reflecting differences in explanatory virtues such as simplicity and elegance. To allow for such cases, their general rule weights the probability of *e* on *h* by the prior probability of *h*; thus $P(h|e) > P(h^*|e)$ if and only if $P(h)P(e|h) > P(h^*)P(e|h^*)$, where $P(e)$, $P(h)$, and $P(h^*)$ are all non-zero. For present purposes, it does not matter whether Bayesians are right to introduce prior probabilities here. The point is that such probabilistic comparisons of hypotheses on the evidence depend on the probabilities of the evidence on the hypotheses. But what has a probability is a proposition; the probability is the probability *that* . . .. At least, that is so when 'probability' has to do with the evidential status of beliefs, as now; if we speak in this connection of the probability of an event, we mean the probability *that it occurred*.[7] We might initially suppose that, in $P(x|y)$, only *x* need be a proposition, but the relation between $P(x|y)$ and $P(y|x)$ means that *y* must be a proposition too; what gives probability must also receive it. Moreover, these probabilities, as measures of degrees of belief warranted by evidence, are idle unless the subject grasps *x* and *y*.

More straightforward uses of evidence also require it to be propositional. In particular, our evidence sometimes rules out some hypotheses by being *inconsistent* with them. For example, the hypothesis that only males have varicose veins is inconsistent with much medical evidence. But only propositions can be inconsistent in the relevant sense. If evidence *e* is inconsistent with an hypothesis *h* in that sense, it must be possible to *deduce* ~*h* from *e*; the premises of a deduction are propositions. Moreover, the subject who deduces ~*h* from *e* must grasp *e*.

Only propositions which we grasp serve the central evidential func-

[7] Objective probabilities, in the sense of chances determined in the natural world independently of our beliefs, are irrelevant here. We sometimes speak, too, of the probability of one property, concept, or predicate conditional on another—for example, the probability of lung cancer conditional on smoking—but the probabilities relevant to the argument are the probabilities of hypotheses, which, unlike properties, concepts, and predicates, have truth-values. That someone has lung cancer is evidence that he smoked; the unattributed property of lung cancer is not by itself evidence of anything.

tions of inference to the best explanation, probabilistic confirmation, and the ruling out of hypotheses. Could non-propositional items count as evidence by serving other central functions of evidence? For example, they might serve as the inputs to a non-inferential process whose outputs were beliefs. But suppose that we are choosing between hypotheses according to which best explains our evidence, or is most probable on our evidence, or is not ruled out by our evidence. The argument so far shows that only propositional evidence would be directly relevant to our choice. Moreover, in choosing between hypotheses in those ways, we can use only propositions which we grasp. In those respects, any evidence other than propositions which we grasp would be impotent. Although evidence may well have central functions additional to those considered above, genuine evidence would make a difference to the serving of the functions considered above, whatever else it made a difference to. Certainly, defences of non-propositional evidence have not been based on an appreciation of its impotence in those respects. Since only propositions we grasp make a difference of the requisite kind, only propositions which we grasp are our evidence.

A positive case for that conclusion has now been given. Nevertheless, perceptual experience is often regarded as a kind of non-propositional evidence. Do the considerations above somehow fail to do it justice? The remainder of this section will rebut objections to the view that our perceptual evidence consists of propositions which we grasp.

Experiences provide evidence; they do not consist of propositions. So much is obvious. But to provide something is not to consist of it. The question is whether experiences provide evidence just by conferring the status of evidence on propositions. On that view, consistent with E = K, the evidence for an hypothesis h consists of propositions $e_1, \ldots, e_n$, which count as evidence for one only because one is undergoing a perceptual experience ε. As a limiting case, h might be e_i. The threatening alternative is that ε can itself be evidence for h, without the mediation of any such $e_1, \ldots, e_n$. Both views permit ε to have a non-propositional, non-conceptual content, but only the latter permits that content to function directly as evidence.

If perceptual evidence consists of propositions, which propositions are they? Consider an example. I am trying to identify a mountain by its shape. I can see that it is pointed; that it is pointed may be part of my evidence for believing that it is not Ben Nevis. However, the proposition that it is pointed does not begin to exhaust my present perceptual evidence. No description of the mountain in words seems to capture the richness of my visual experience of its irregular shape. But it does not follow that my evidence is non-propositional. If I want to convey my evidence, I might point and say 'It is that shape'.[8] Of course, the mere linguistic meaning of the sentence type 'It is that shape' does not convey my evidence, for it is independent of the reference of 'that shape' in a particular context of utterance. Only by using the sentence in an appropriate context do I express the proposition at issue. My token of 'that shape' still expresses a constituent of that proposition, even if you cannot grasp that constituent without having a complex visual experience with a structure quite different from the constituent structure of the proposition. The proposition that the mountain is that shape is contingent; it could have been another shape. The proposition is also known a posteriori; I do not know a priori that I am not including the tip of another mountain behind in the profile. But in ordinary circumstances I can know that the mountain is that shape, and a fortiori grasp the proposition that it is, when 'that shape' does not refer to an absolutely specific shape. Of course, I cannot see *exactly* what shape the mountain is; I can only see roughly what profile it presents to me, and cannot see round the back. That shape must be unspecific enough to give my knowledge that the mountain is that shape an adequate margin for error in the sense of Chapter 5.[9] The knowledge that the mountain is that shape is obtainable in other contexts; you can have it too, and we can retain it in memory. Properties other than shape are similar in those respects.

In unfavourable circumstances, one fails to gain perceptual knowledge, perhaps because things are not the way they appear to be. One does not know that things are that way, and E = K excludes the proposition that they are as evidence. Nevertheless, one still has perceptual evidence, even if the propositions it supports are false. True propositions can make a false proposition probable, as when someone is skilfully framed for a crime of which she is innocent. If perceptual evidence in the case of illusions consists of true propositions, what are they? The obvious answer is: the proposition that things appear to be that way. The mountain appears to be that shape. Of course, unless one has reason to suspect that circumstances are unfavourable, one may not consider the cautious proposition that things appear to be that way; one

[8] See McDowell 1994: 56–9 for such a proposal. It is not being used here to deny that perceptual experience has non-conceptual content (see Peacocke 1992: 84). Christensen 1992: 545 discusses such beliefs in a Bayesian context, asking whether they can 'connect with other beliefs in the way that would be necessary for them to fulfill their intended evidential role'. The connections will not be purely syntactic, but fifty years of confirmation theory have shown that confirmation is not a purely syntactic matter.

[9] The unspecificity makes the present proposal closer to that of McDowell 1994: 170–1 than to that of Peacocke 1992: 83–4.

may consider only the unqualified proposition that they really are that way. But it does not follow that one does not know that things appear to be that way, for one knows many propositions without considering them. When one is walking, one normally knows that one is walking, without considering the proposition. Knowing is a state, not an activity. In that sense, one can know without consideration that things appear to be some way. When I believe falsely that circumstances are favourable, I believe falsely that I am gaining perceptual knowledge about the environment, and therefore that my evidence includes those propositions believed to be known. But our fallibility in identifying our evidence is nothing new, and my actual evidence may justify my false beliefs about my evidence.

In order to grasp the proposition that things appear to be some way, one must grasp the property of appearing, on the assumption that the semantically significant constituents of a sentence express constituents of the proposition expressed by the whole sentence. Although one's grasp of the property of appearing may be inarticulate, one must have some inkling of the distinction between appearance and reality. For instance, one should be willing in appropriate circumstances to give up the belief that things were that way while retaining the belief that they appeared to be that way. In the absence of such dispositions, it is implausible to attribute the qualified belief that things appear to be that way rather than the unqualified belief that they are that way. Perhaps some young children and animals have beliefs and perceptual experiences without even implicitly grasping the property of appearance. Suppose that such a simple creature is given a drug which causes the hallucinatory appearance that there is food ahead; as a result, it comes to believe falsely that there is food ahead. Does it have any evidence for that belief? According to E = K, its evidence cannot be that things appear some way, for it cannot grasp that proposition. Perhaps it knows that the situation is *like* one in which there is food ahead, where the property of likeness covers both likeness in appearance and other kinds of likeness indifferently, so that grasp of the property of likeness does not require grasp of the property of appearing. If the creature does not even know that the situation is like one in which there is food ahead, then we can plausibly deny that it has perceptual evidence that there is food ahead. It does not recognize the features of its perceptual experience which, if recognized, would provide it with evidence. *We* can use the proposition that there appears to be food ahead as evidence, but the simple creature cannot. Although the hallucinatory appearance causes a belief, that causal relation is not an evidential one.

Very simple creatures grasp no properties or propositions and have no beliefs or knowledge. It is sometimes even argued—not very plausibly—that any creature which lacks the distinction between appearance and reality is in this predicament. Simple creatures have no evidence, for they have no degrees of belief, and degrees of belief are what evidence justifies.

S can use as evidence only propositions which S grasps. Since S can use S's evidence as evidence, only propositions which S grasps are S's evidence. What has not yet been argued is that those propositions count as evidence by being known.

9.6 PROPOSITIONAL EVIDENCE AS KNOWLEDGE

Why should all propositional evidence be knowledge? The thesis is that if S's evidence includes a proposition *e*, then S knows *e*. If I do not know that the mountain is that shape, then that it is that shape is not part of my evidence. As in the previous section, the argument is from the function of evidence.[10] Indeed, the thesis draws support from the role of evidence cited there, in inference to the best explanation, probabilistic reasoning, and the exclusion of hypotheses. When we prefer an hypothesis *h* to an hypothesis *h** because *h* explains our evidence *e* better than *h** does, we are standardly assuming *e* to be known; if we do not know *e*, why should *h*'s capacity to explain *e* confirm *h* for us? It is likewise hard to see why the probability of *h* on *e* should regulate our degree of belief in *h* unless we know *e*. Again, an incompatibility between *h* and *e* does not rule out *h* unless *e* is known. But it is prudent to consider the matter more carefully.

Suppose that balls are drawn from a bag, with replacement. In order to avoid issues about the present truth-values of statements about the future, assume that someone else has already made the draws; I watch them on film. For a suitable number *n*, the following situation can arise. I have seen draws 1 to *n*; each was red (produced a red ball). I have not yet seen draw *n*+1. I reason probabilistically, and form a justified belief that draw *n*+1 was red too. My belief is in fact true. But I do not know that draw *n*+1 was red. Consider two false hypotheses:

h: Draws 1 to *n* were red; draw *n*+1 was black.

*h**: Draw 1 was black; draws 2 to *n*+1 were red.

[10] For linguistic arguments that if S's reason or justification is *e* then S knows *e*, see Unger 1975: 206–14. See also Hyman 1999.

It is natural to say that *h* is consistent with my evidence and that *h** is not. In particular, it is consistent with my evidence that draw *n*+1 was black; it is not consistent with my evidence that draw 1 was black. Thus my evidence does not include the proposition that draw *n*+1 was red. Why not? After all, by hypothesis I have a justified true belief that it was red. The obvious answer is that I do not *know* that draw *n*+1 was red; the unsatisfied necessary condition for evidence is knowledge. An alternative answer is that I have not *observed* that draw *n*+1 was red. That is equally good for the purposes of this section (although not for those of the next), for observing the truth of *e* includes *e* in my evidence only by letting me know *e*. If I observe the truth of *e* and then forget all about it, my evidence no longer includes *e*. It is hard to see how evidence could discriminate between hypotheses in the way we want it to if it did not have to be known.

If evidence required only justified true belief, or some other good cognitive status short of knowledge, then a critical mass of evidence could set off a kind of chain reaction. Our known evidence justifies belief in various true hypotheses; they would count as evidence too, so this larger evidence set would justify belief in still more true hypotheses, which would in turn count as further evidence . . .. The result would be very different from our present conception of evidence.

That propositional evidence is knowledge entails that propositional evidence is true. That is intuitively plausible; if one's evidence included falsehoods, it would rule out some truths, by being inconsistent with them. One's evidence may make some truths improbable, but it should not exclude any outright. Although we may treat false propositions as evidence, it does not follow that they are evidence. No true proposition is inconsistent with my evidence, although I may think that it is. If *e* is evidence for *h*, then *e* is true. There is no suggestion, of course, that if *e* is evidence for *h* then *h* is true. For example, that the ground is wet is evidence that it rained last night only if the ground is wet—even if it did not rain last night. If *e* is not true, then at most a counterfactual holds: if *e* had been true, *e* would have been evidence for *h*.[11] If the convincing but lying witness says that the accused was asleep at the time of the murder, then it is part of the evidence for the innocence of the accused that the witness *said* that he was asleep then. It is not part of the

[11] Stampe 1987: 337 takes a similar view of reasons. Millar 1991: 65 says that we ordinarily think of evidence as consisting in facts; that is also how we ordinarily think of the objects of knowledge. If facts are distinguished from true propositions, then the arguments of this chapter can be adjusted accordingly, but the individuation of facts must be reconciled with the individuation of knowledge and evidence. Presumably, the fact that Hesperus is bright is the fact that Phosphorus is bright. See also section 1.5.

evidence for his innocence that he was asleep, for it is consistent with the evidence that he was not. The rival view, that a false proposition can become evidence through a sufficient appearance of truth, gains most of its appeal from the assumption, disposed of in Chapter 8 and section 9.3, that we must have an infallible way of identifying our evidence.

Once it is granted that all propositional evidence is true—and therefore, by the previous section, that all evidence consists of true propositions—adjusting our beliefs to the evidence has an obvious point. It is a way of adjusting them to the truth. Although true evidence can still support false conclusions, it will tend to support truths. The maxim 'Proportion your belief to your evidence' requires more than the mere internal coherence of one's belief system; it does so because evidence must be true. Even if an internally coherent belief system cannot be wholly false, a given belief system with a given degree of internal coherence can be better or worse proportioned to the evidence, depending on what the evidence is. But, equally, the evidence is not a wholly external standard, if it is known.

Another consequence of the claim that propositional evidence is knowledge is that propositional evidence is believed—at least, if knowledge entails belief, which is granted here (see section 1.5). The case of perception may seem to suggest that propositional evidence is not always believed. In conformity with the previous section, a piece of perceptual evidence is, for example, a proposition *e* that things are *that* way. According to E = K, my evidence includes *e* because I know that things are that way. But, a critic may suggest, that does not go back far enough; my evidence includes *e* because it is perceptually apparent to me that things are that way, whether or not I believe that they are that way. Even if I do believe *e*, my evidence included *e* before I came to believe it; according to the critic, I came to believe it because it was perceptually apparent. If 'It is perceptually apparent that A' entails 'A', then the critic's view allows that evidential propositions are always true; what it denies is that they are always believed, and therefore that they are always known.

If my evidence includes a proposition *e*, then I grasp *e*, by section 9.5. Thus, if I fail to believe *e*, my problem is not conceptual incapacity. Perhaps I have simply not had time to form the belief; perhaps I suspect, for good or bad reasons, that I am the victim of an illusion. We can ask the critic whether, for my evidence to include *e*, I must at least be *in a position* to know *e*? If so, then the critic's view does not differ radically from E = K. Given E = K, the evidence in my actual possession consists of the propositions which I know, but there is also the evidence in my potential possession, consisting of the propositions which I am in a position to know. The critic takes my evidence to be the evidence in my potential possession, not just the evidence in my actual possession. To bring out the difference between that view and E = K, suppose that I am in a position to know any one of the propositions $p_1, \ldots, p_n$ without being in a position to know all of them; there is a limit to how many things I can attend to at once. Suppose that in fact I know p_1 and do not know $p_2, \ldots, p_n$. According to E = K, my evidence includes only p_1; according to the critic, it includes $p_1, \ldots, p_n$. Let q be a proposition which is highly probable given $p_1, \ldots, p_n$ together, but highly improbable given any proper subset of them; the rest of my evidence is irrelevant to q. According to E = K, q is highly improbable on my evidence. According to the critic, q is highly probable on my evidence. E = K gives the more plausible verdict, because the high probability of q depends on an evidence set to which as a whole I have no access.

The contrast with E = K is more radical if the critic allows my evidence to include *e* even when I am not in a position to know *e*. For example, it is perceptually apparent to me that it is snowing; I am not hallucinating; but since I know that I have taken a drug which has a 50 per cent chance of causing me to hallucinate, I am not in a position to know that it is snowing. According to the radical critic, my evidence nevertheless includes the proposition that it is snowing, because it is perceptually apparent to me that it is snowing; thus my evidence is inconsistent with the hypothesis that I am hallucinating and it is not snowing, even though, for all I am in a position to know, that hypothesis is true. According to E = K, my evidence includes at best the proposition that it appear to be snowing. Surely, if I proportion my belief to my evidence, I shall not dismiss the hypothesis that I am hallucinating and it is not snowing. E = K gives the better verdict. Perceptual cases do not show that we sometimes fail to believe our evidence.

A truth does not become evidence merely by being believed, or even by being justifiably believed, as the example of the proposition that draw $n+1$ was red showed above. Nothing short of knowledge will do. But is even knowledge enough?

9.7 KNOWLEDGE AS EVIDENCE

Any restriction on what counts as evidence should be well-motivated by the function of evidence. By sections 9.5 and 9.6, one's evidence includes only propositions which one knows. If, when assessing an hypothesis, one knows something *e* which bears on its truth, should not *e* be part of one's evidence? Would it not violate the total evidence condition to do otherwise? This section examines attempts to justify some further restriction on evidence, and finds them wanting.

One's knowledge is held together by a tangle of evidential interconnections. For example, my knowledge that Henry V died in 1422 is evidentially related to my knowledge that various books say that he died in 1422. Much of one's knowledge is redundant, in the sense that the proposition known is a logical consequence of other known propositions. Perhaps each proposition which I know is redundant in that sense. If all knowledge is evidence, the evidential interconnectedness and redundancy is internal to one's evidence. The redundancy itself is harmless; it does not make the evidence support the wrong hypotheses. The concern is rather that if all one's knowledge is treated as a single body of evidence, its internal evidential interconnections will be obliterated, and therefore that such an account would falsify the nature of our knowledge.

The alternative, presumably, is for evidence to be self-evident, consisting of epistemically self-sufficient nuggets of information. That is an implausibly atomistic picture of evidence, but it constitutes a challenge to explain how there can be evidential interconnections within a single body of evidence. Section 9.2 provides the basis for such an explanation. According to EV, when e is evidence for an hypothesis h for one, one's evidence includes e, and e raises the probability of h, which requires the probability of e on the relevant distribution to be less than 1. Thus EV already permits one proposition in one's evidence to be evidence for another in a non-trivial way. The internal evidential interconnections are not obliterated.

If all knowledge is evidence, then EV in section 9.2 does have the effect of making evidential interconnections within one's knowledge symmetric. For $P(p \mid q) > P(p)$ if and only if $P(p \wedge q) > P(p)P(q)$; since the latter condition is symmetric in p and q, $P(p \mid q) > P(p)$ if and only if $P(q \mid p) > P(q)$. Thus, given that S's evidence includes both p and q, p is evidence for q for S if and only if q is evidence for p for S by EV. Consequently, given that one knows p and q and that all knowledge is evidence, EV implies that if p is evidence for q for one then q is evidence for p for one. We could avoid this result by modifying EV. For example, we could stipulate that e is evidence for h for S only if S's belief in e does not essentially depend on inference from h. But it might be neater to retain EV unmodified and say that e is *independent* evidence for h for S only if S's belief in e does not essentially depend on inference from h. Since the focus of this discussion is not on the evidence-for relation, we shall not pursue these options further.

The claim that all knowledge is evidence faces another sort of objection. Very little (if any) of what we know is indubitable. Therefore, if all knowledge is evidence, much of our evidence is dubitable. We are uneasy with the idea of uncertain evidence. Is this just the old Cartesian prejudice that only unshakable foundations will do? The worry cannot be so easily dismissed. It takes a particularly sharp form in a Bayesian context. The standard way of accommodating new evidence *e* is by conditionalizing on it. The new unconditional probability of a proposition is its old probability conditional on *e* (where the old probability of *e* was non-zero); $P_{new}(h) = P_{old}(h \mid e)$. In particular, $P_{new}(e) = P_{old}(e \mid e) = 1$. These probability distributions should be distinguished from P in EV above, for both P_{old} and P_{new} are supposed to incorporate all of one's evidence at the relevant times; whereas it was observed that P must incorporate only a proper part of one's evidence. Now if the old probability of *h* was 1, so is its new probability; for if $P_{old}(h) = 1$ then $P_{old}(h \wedge e) = P_{old}(e)$. Since the new probability of *e* is 1, it will remain 1 under any series of conditionalizations on further propositions. Thus once a proposition is conditionalized on as evidence, it acquires probability 1, and retains it no matter what further evidence is conditionalized on. But most of our knowledge has no such status. Further evidence could undermine it.[12]

Here is an example. I put exactly one red ball and one black ball into an empty bag, and will make draws with replacement. Let *h* be the proposition that I put a black ball into the bag, and *e* the proposition that the first ten thousand draws are all red. I know *h* by a standard combination of perception and memory, because I saw that the ball was black as I put it into the bag a moment ago. Nevertheless, if after ten thousand draws I learn *e*, I shall have ceased to know *h*, because the evidence which I shall then have will make it too likely that I was somehow confused about the colours of the balls. Of course, what I know now is true, and so will never be discovered to be false, but it does not follow that there will never be misleading future evidence against it. My present knowledge is consistent with *e*; on simple assumptions, *e* has a probability of $1/2^{10,000}$ on my present evidence. If I subsequently learn *e*, the probability of *h* on that future evidence will be less than 1. But if conditionalization on subsequent evidence will give *h* a probability less than 1, then the present probability of *h* is less than 1, so *h* is not part of my present evidence. The problem is general: if misleading future evidence of positive probability can undermine my knowledge that I put a

[12] The use of Jeffrey conditionalization to avoid the problem is criticized in section 10.2.

black ball into the bag, it can undermine most of my present knowledge. It looks as though *h* should count as part of my present evidence, and therefore receive probability 1, only if *h* is bound to be a rational belief for me in the future, come what may. Few propositions will pass that test. Indeed, not even *e* passes the test, for later evidence may make it rational for me to believe that I had misremembered the outcome of the first ten thousand draws; several eyewitnesses may insist that I was misremembering; but since uncertainty about *e* does not make *h* certain, it does not rehabilitate *h* as evidence. By this line of argument, either we know very little, or very little of our knowledge is evidence.

What empirical propositions qualify as evidence by the proposed test that their probability should never subsequently slip below 1? One might suppose that the best candidates would be propositions about the present—traditionally, propositions about the subject's present mental states. Since the test requires evidence to remain certain as time passes, in order for a proposition about the present to be evidence, it must remain certain long after the time it is about has passed. But even if it is absolutely certain for me today that I seem to see a blue patch, it will not be absolutely certain for me tomorrow that I seemed to see a blue patch today; my memories will not be beyond question. It would only exacerbate the problem to individuate propositions so that the present-tensed sentence 'I seem to see a blue patch' expressed the same proposition at different times, for even if such a proposition is certain and so true now, it will be false and so uncertain in the future. It is hard to see what empirical propositions would qualify as evidence by the proposed test. Thus the very possibility of learning from experience is threatened.

The model assumes that probabilities change only by conditionalization on new evidence. This is to assume that evidence can be added but not subtracted over time. The assumption is obviously false in practice, because we sometimes forget. But even if the model is applied to elephants, idealized subjects who never forget, the assumption that evidence cannot be lost is implausible. On any reasonable theory of evidence, an empirical proposition which now counts as evidence can subsequently lose its status as evidence without any forgetting, if future evidence casts sufficient doubt on it. Given E = K, this process is the undermining of knowledge. The next chapter develops a more liberal model within a broadly Bayesian framework in which evidence can be lost as well as gained. If today's evidence is not evidence tomorrow, its probability tomorrow can be less than 1. The requirement that the probability of present evidence should never slip below 1 in the future was just an artefact of an overly restrictive model of updating.

One could have a model of the same structure on which only knowledge of the present, or observational knowledge, counts as evidence.[13] But once it is recognized that evidence is not obliged to meet unusual standards of certainty, such restrictions on evidence look ad hoc. Although knowledge of the present or observational knowledge may be easier to obtain than some other kinds of knowledge, that is no reason against counting other kinds of knowledge as evidence, when we obtain them. For example, our evidence for a mathematical conjecture may consist of mathematical knowledge. If we believe that we know *p*, we shall be disposed to use *p* in the ways in which we use evidence. If our belief is true, we are right to use *p* in those ways. It does not matter what kind of proposition *p* is; as Austin said, '*any* kind of statement could state evidence for *any* other kind, if the circumstances were appropriate' (1962: 116). All knowledge is evidence.

9.8 NON-PRAGMATIC JUSTIFICATION

The present case for E = K is now complete. If evidence is what justifies belief, then knowledge is what justifies belief. But is all justified belief justified by evidence? Why cannot experience itself, or practical utility, justify belief? Why cannot belief sometimes *be* justified without being justified *by* anything at all?

The *pragmatic* justification of belief need not be by evidence. Without any evidence at all, someone believes that her child somehow survived an air crash, and will one day return to her. The belief is the only thing which keeps her going; without it, she would kill herself. Perhaps it is on balance a good thing that she has the belief, and in that sense the belief is justified. But this is not the sense of 'justified' in which justified belief appeared to have marginalized knowledge within epistemology. Could belief be *epistemically* justified except by evidence? Epistemic justification aims at truth in a sense—admittedly hard to define—in which pragmatic justification does not. It is far from obvious that any belief is justified in the truth-directed sense without being justified by evidence. It appears otherwise when evidence is conceived too

[13] For Maher, '*E* is evidence iff *E* is known directly by experience' (1996: 158; he relativizes 'directly' to a set of propositions, 160–1). On his view, if I know *e* (e.g. that a substance s dissolved when placed in water), and deduce *h* (e.g. that s is soluble), thereby coming to know *h*, then *e* but not *h* is evidence for me (1996: 158). On the present view, both *e* and *h* are evidence for me, but *h* is not independent evidence for *e*, since in the circumstances I believe *e* by inference from *h*. If someone now tells me that s is salt, *h* may become independent evidence for *e*.

narrowly, for then the evidence looks too scanty to justify all the beliefs which are in fact justified. But if anything we know can be evidence to anchor a chain of justification, as E = K implies, then evidence plausibly suffices for all truth-directed justification. An epistemically justified belief which falls short of knowledge must be epistemically justified by something; whatever justifies it is evidence. An epistemically justified belief which does not fall short of knowledge is itself evidence, by E = K. If we are aiming at the truth, we should proportion our belief to the evidence.

E = K supports the plausible equation of truth-directed justification with justification by evidence, and therefore with justification by knowledge. On this view, if truth-directed justification is central to epistemology, so too is knowledge.

We can suggest something more radical. Belief does not aim merely at truth; it aims at knowledge. The more it is justified by knowledge, the closer it comes to knowledge itself. If evidence and knowledge are one, then the more a belief is justified by evidence, the closer it comes to its aim. The next two chapters will help to make those suggestions good.

10
Evidential Probability

10.1 VAGUE PROBABILITY

When we give evidence for our theories, the propositions which we cite as evidence are themselves uncertain. Probabilistic theories of evidence have notorious difficulty in accommodating that obvious fact, as section 9.7 noted. This chapter embeds the fact in a probabilistic theory of evidence. The analysis of uncertainty leads naturally to a simple theory of higher-order probabilities. The first step is to focus on the relevant notion of probability.

Given a scientific hypothesis *h*, we can intelligibly ask: how probable is *h* on present evidence? We are asking how much the evidence tells for or against the hypothesis. We are not asking what objective physical chance or frequency of truth *h* has. A proposed law of nature may be quite improbable on present evidence even though its objective chance of truth is 1. That is quite consistent with the obvious point that the evidence bearing on *h* may include evidence about objective chances or frequencies. Equally, in asking how probable *h* is on present evidence, we are not asking about anyone's actual degree of belief in *h*. Present evidence may tell strongly against *h*, even though everyone is irrationally certain of *h*. We will refer to degrees of belief as *credences*; for example, one's prior credence in the proposition that the fair coin will come up heads is normally 1/2; thus credences are *not* the degrees of outright belief discussed in section 4.4.

Is the probability of *h* on our evidence the credence which a perfectly rational being with our evidence would give to *h*? That suggestion comes closer to what is intended, but not close enough. It fails in the way in which counterfactual analyses usually fail, by ignoring side-effects of the conditional's antecedent on the truth-value of the analysandum (Shope 1978). For example, to say that the hypothesis that there are no perfectly rational beings is very probable on our evidence is not to say that a perfectly rational being with our evidence would be very confident that there were no perfectly rational beings. To make the point more carefully, let *p* be a logical truth (a proposition expressed by a logically true sentence) such that in this imperfect world it is very probable on our evidence that no one has great credence in *p*. There are such logical truths, although in the nature of the case we cannot be confident that we have identified an example. For all we know, they include the proposition that Goldbach's Conjecture is a theorem of first-order Peano Arithmetic (appropriately formalized). Of course, it is not highly probable on our evidence that no one will *ever* give high credence to the proposition that Goldbach's Conjecture is a theorem of first-order Peano arithmetic; we can eternalize the example, if we like, by imagining good evidence that nuclear war is about to end all intelligent life. Let *h* be the hypothesis that no one has great credence in *p*. By assumption, *h* is very probable on our evidence. On the view in question, a perfectly rational being with our evidence would therefore have great credence in *h*. Since *p* is a logical truth, *h* is logically equivalent to the conjunction $p \wedge h$; since a perfectly rational being would have the same credence in logically equivalent hypotheses, it would have great credence in $p \wedge h$. But that is absurd, for $p \wedge h$ is of the Moore-paradoxical form 'A and no one has great credence in the proposition that A'; to have great credence in $p \wedge h$ would therefore be self-defeating and irrational. One can have great credence in a true proposition of that form only by irrationally having greater credence in the conjunction than in its first conjunct. Thus the probability of a hypothesis on our evidence does not always coincide with the credence which a perfectly rational being with our evidence would have in it.

Presumably, a perfectly rational being must give great credence to *p*, be aware of doing so, and therefore give little credence to *h* and so to $p \wedge h$; but then its evidence about its own states would be different from ours. If so, the hypothesis of a perfectly rational being with our evidence is impossible. There is no such thing as *the* credence which a perfectly rational being with our evidence would have in a given proposition. It can be argued that the subjective Bayesian conception of perfect rationality entails perfect accuracy about one's own credences (Milne 1991).

We therefore cannot use decision theory as a guide to evidential probability. Suppose, for example, that anyone whose credences have distribution P is vulnerable to a Dutch Book, a complex bet on which they lose money no matter what the outcome. It may follow that the credences of a perfectly rational being would not have distribution P, if a perfectly rational being would not be vulnerable to a Dutch Book, but it would be fallacious to conclude that probabilities on our evidence do not have distribution P, for those probabilities need not coincide with the hypothetical credences of a perfectly rational being. Perhaps only an imperfectly rational being could have exactly our evidence, which includes our evidence about ourselves. The irrationality of distributing credences according to the probabilities on one's evidence may simply reflect one's limited rationality, as reflected in one's evidence. But it would be foolish to respond by confining evidential probability to evidence sets which could be the total evidence possessed by a perfectly rational creature. That would largely void the notion of interest; we care about probabilities on *our* evidence.

For all that has been said, any agent with credences which fail to satisfy subjective Bayesian constraints may be *eo ipso* subject to rational criticism. This would apply in particular to the agent's beliefs *about* probabilities on its evidence. But it would apply equally to the agent's beliefs about objective physical chances, or anything else. Just as it implies nothing specific about objective physical chances, so it implies nothing specific about probabilities on evidence.

What, then, *are* probabilities on evidence? We should resist demands for an operational definition; such demands are as damaging in the philosophy of science as they are in science itself. To require mathematicians to give a precise definition of 'set' would be to abolish set theory. Sometimes the best policy is to go ahead and theorize with a vague but powerful notion. One's original intuitive understanding becomes refined as a result, although rarely to the point of a definition in precise pretheoretic terms. That policy will be pursued here. The discussion will assume an initial probability distribution P. P does not represent actual or hypothetical credences. Rather, P measures something like the intrinsic plausibility of hypotheses prior to investigation; this notion of intrinsic plausibility can vary in extension between contexts. P will be assumed to satisfy a standard set of axioms for the probability calculus: $P(p)$ is a non-negative real number for every proposition p; $P(p) = 1$ whenever p is a logical truth; $P(p \vee q) = P(p) + P(q)$ whenever p is inconsistent with q. If $P(q) > 0$, then the conditional probability of p on q, $P(p \mid q)$, is defined as $P(p \wedge q)/P(q)$. $P(p)$ is taken to be defined for all propositions; the standard objection that the subject may never have considered p is irrelevant to the non-subjective probability P. But P is *not* assumed to be syntactically definable. Carnap's programme of inductive logic is moribund. The difference between green and grue is not a formal one.

Consider an analogy. The concept of *possibility* is vague and cannot be defined syntactically. But that does not show that it is spurious. In fact, it is indispensable. Moreover, we know some sharp structural constraints on it: for example, that a disjunction is possible if and only if at least one of its disjuncts is possible. The present suggestion is that probability is in the same boat as possibility, and not too much the worse for that.

On the view to be defended here, the probability of a hypothesis *h* on total evidence *e* is P(*h* | *e*). The last chapter gave an account of when a proposition *e* constitutes one's total evidence. The best that evidence can do for a hypothesis is to entail it (so P(*h* | *e*) = 1); the worst that evidence can do is to be inconsistent with it (so P(*h* | *e*) = 0). Between those extremes, the initial probability distribution provides a continuum of intermediate cases, in which the evidence comes more or less close to requiring or ruling out the hypothesis.

The axioms entail that logically equivalent propositions have the same probability on given evidence. The reason is not that a perfectly rational being would have the same credence in them, for the irrelevance of such beings to evidential probability has already been noted. The axioms are *not* idealizations, false in the real world. Rather, they show what kind of thing we are choosing to study. We are using a notion of probability which (like the notion of incompatibility) is insensitive to differences between logically equivalent propositions. We thereby gain mathematical power and simplicity at the loss of some descriptive detail (for example, in the epistemology of mathematics): a familiar bargain.

The characterization of the prior distribution for evidential probability is blatantly vague. If that seems to disadvantage it with respect to subjective Bayesian credences, which can be more precisely defined in terms of consistent betting behaviour, the contrast in precision disappears in epistemological applications. Given a finite body of evidence *e*, almost any posterior distribution results from a sufficiently eccentric prior distribution by Bayesian updating on *e*. Theorems on the 'washing out' of differences between priors by updating on evidence apply only 'in the limit'; they tell us nothing about where we are now (Earman 1992: 137–61 has a sophisticated discussion). Successful Bayesian treatments of specific epistemological problems (for example, Hempel's paradox of the ravens) assume that subjects have 'reasonable' prior distributions. We judge a prior distribution reasonable if it complies with our intuitions about the intrinsic plausibility of hypotheses. This is the same sort of vagueness as infects the present approach, if slightly better hidden.

One strength of Bayesianism is that the mathematical structure of the probability calculus allows it to make illuminating distinctions which other approaches miss and provide a qualitatively fine-grained analysis of epistemological problems, given assumptions about all reasonable prior distributions. That strength is common to subjective and objective Bayesianism, for it depends on the structure of the probability calculus. On the present approach, which can be regarded as a form of objective

Bayesianism, the axioms of probability theory embody substantive claims, as the axioms of set theory do. For example, the restriction of probabilities to real numbers limits the number of gradations in probability to the cardinality of the continuum. Just as the axioms of set theory refine our understanding of sets without reducing to implicit definitions of 'set', so the axioms of probability theory refine our understanding of evidential probability without reducing to implicit definitions of 'evidential probability'.

The remarks above are not intended to smother all doubts about the initial probability distribution. Their aim is to justify the procedure of tentatively postulating such a distribution, in order to see what use can be made of it in developing a theory of evidential probability. That is the focus of this chapter.[1]

10.2 UNCERTAIN EVIDENCE

Suppose that evidential probabilities are indeed probabilities conditional on one's evidence. Then, trivially, the evidence itself has evidential probability 1. $P(e|e) = 1$ whenever it is defined. Does this require evidence to be absolutely certain? If so, how can evidential probabilities fit into a non-Cartesian epistemology? Section 9.7 gave the problem a preliminary discussion. Let us now consider it more thoroughly.

Section 9.5 defended the assumption that evidence is propositional. Since the approach in this chapter identifies evidential probabilities with probabilities conditional on the evidence, it is in any case committed to treating evidence as propositional. $P(h | e) = P(h \wedge e)/P(e)$; this equation makes sense only if the evidence e is propositional. We therefore cannot avoid attributing evidential probability 1 to the evidence by denying that evidence is propositional, for then evidential probabilities would be undefined.

We should question the association between evidential probability 1 and absolute certainty. For subjective Bayesians, probability 1 is the highest possible degree of belief, which presumably is absolute certainty. If one's credence in p is 1, one should be willing to accept a bet on which one gains a penny if p is true and is tortured horribly to death if p is false. Few propositions pass that test. Surely complex logical truths do not, even though the probability axioms assign them probability 1.

[1] No attempt will be made to survey the non-Bayesian theories of evidential probability in the literature. See e.g. Kyburg 1974 and Plantinga 1993.

But since evidential probabilities are not actual or counterfactual credences, why should evidential probability 1 entail absolute certainty?

There is a further link between probability 1 and certainty. Bayesian accounts of learning from experience give a significance to probability 1 which does not depend on any identification of probabilities with actual or counterfactual credences. Suppose that the new evidence gained on some occasion is *e*. On the standard Bayesian account of this simple case, probabilities should be updated by *conditionalization* on *e*. The updated unconditional probability of *p* is its previous probability conditional on *e*:

$$\text{BCOND} \quad P_{new}(p) = P_{old}(p|e) = P_{old}(p \wedge e)/P_{old}(e) \qquad (P_{old}(e) \neq 0)$$

We can interpret BCOND as a claim about evidential probabilities. Note that P_{old} is not absolutely prior probability P, but probability on all the evidence gained prior to *e*. Suppose further, as Bayesians often do, that such conditionalization is the only form of updating which the probabilities undergo. By BCOND, $P_{new}(e) = 1$. When P_{new} is updated to P_{vnew} by conditionalization on still newer evidence *f*, $P_{vnew}(e) = P_{new}(e|f) = P_{new}(e \wedge f)/P_{new}(f) = 1$ whenever conditionalization on *f* is defined. Thus *e* will retain probability 1 through all further conditionalizations. Since no other form of updating is contemplated, *e* will retain probability 1. Once a proposition has been evidence, its status is as good as evidence ever after; probability 1 is a lifetime's commitment. On this model of updating, when a proposition becomes evidence it acquires an epistemically privileged feature which it cannot subsequently lose. How can that be? Surely any proposition learnt from experience can in principle be epistemically undermined by further experience.

What propositions could attain that unassailable epistemic status? Science treats as evidence propositions such as 'Thirteen of the twenty rats injected with the drug died within twenty-four hours'; one may discover tomorrow that a disaffected laboratory technician had substituted dead rats for living ones. The Cartesian move is to find certainty in propositions about one's own current mental state ('I seem to see a dead rat'; 'My current degree of belief that thirteen of the twenty rats died is 0.97'). Arguably, we are fallible even about our own current mental states (see Chapters 4 and 8). But even if that point is waived, and we are assumed to be infallible about a mental state when we are in it, we do not remain infallible about it later. However certain I am today of the proposition which I now express by the sentence 'I seem to see a dead rat', I may be uncertain tomorrow of the same proposition, then expressed by the sentence 'Yesterday I seemed to see a dead rat'. I can wonder whether I really remember seeming to see a dead rat, or only imagine it. Perhaps 'I seem to see a dead rat' (uttered by me today) and 'Yesterday I seemed to see a dead rat' (uttered by me tomorrow) do not express exactly the same proposition. But if I can think tomorrow the proposition expressed by 'I seem to see a dead rat' (uttered by me today), then that proposition can become uncertain for me. If I cannot even think it tomorrow, then the problem is even worse, because I *cannot* retain my evidence. We are uncontroversially fallible about our own past mental states. We are likewise fallible about the mental states of others. You can doubt whether I seem to myself to see a dead rat. Even if I tell you that I seem to myself to see one, you may wonder whether I am lying. Yet science relies on intersubjectively available evidence. Even Bayesian epistemologists assume that evidence is intersubjectively available. Consider, for instance, the arguments that individual differences between prior probability distributions are 'washed out' in the long run by conditionalization on accumulating evidence. They typically assume that different individuals are conditionalizing on the *same* evidence. If we start with different prior probabilities, and I conditionalize on evidence about my mental state while you conditionalize on evidence about your mental state, then our posterior probabilities need not converge.

In some cases it can be shown that, although our evidence is different, our beliefs will almost certainly converge on each other because they will almost certainly converge on the truth. For example, if a bag contains ten red or black balls, and we take it in turns to draw a ball with replacement, each observing our own draws and not the other's, and conditionalizing on the results, our posterior probabilities for the number of balls of each colour will almost certainly converge to the same values, even if our prior probabilities are quite different, provided that we both assign non-zero prior probabilities to all eleven possibilities. But even this assumes that our evidence consists of true propositions about the results of the draws, not propositions about our mental states. Where does that assumption come from, on a subjective Bayesian view?

The point generalizes. It is tempting to make a proposition *p* certain for a subject S at a time *t* by attributing a special authority to S's belief at *t* in *p*. But then belief in *p* by other subjects or at other times has a special lack of authority, because it is trumped by S's belief at *t*. For example, to the extent to which eyewitness reports of an event have a special status, non-eyewitness reports are vulnerable to being overturned by them. Thus it is hard to see how *any* empirical proposition could have the intertemporal and intersubjective certainty which the conditionalization account demands of evidence.

The standard response is to generalize Bayesian conditionalization to

Jeffrey conditionalization (probability kinematics). For a proposition p, in Bayesian conditionalization on e ($0 < P_{old}(e) < 1$):

$$\text{(i) } P_{old}(p) = P_{old}(e)P_{old}(p \mid e) + P_{old}(\sim e)P_{old}(p \mid \sim e)$$

$$\text{(ii) } P_{new}(p) = P_{new}(e)P_{old}(p \mid e) + P_{new}(\sim e)P_{old}(p \mid \sim e)$$

For BCOND, the weights $P_{new}(e)$ and $P_{new}(\sim e)$ in (ii) are 1 and 0 respectively. Probabilities conditional on e are unchanged ($P_{new}(p|e) = P_{old}(p|e)$). What has changed is their weight in determining unconditional probabilities; it has increased from $P_{old}(e)$ to 1. But when experience makes e more probable without making it certain, Jeffrey conditionalization allows us to retain (ii) ((i) is automatic) and make $P_{new}(e)$ larger than $P_{old}(e)$ without making it 1. This increases the weight of probabilities conditional on e at the expense of probabilities conditional on $\sim e$, while giving some weight to both. More generally, experience may cause us to redistribute probability amongst various possibilities, whilst leaving probabilities conditional on those possibilities fixed. Let $\{e_1, \ldots, e_n\}$ be a partition (that is, as a matter of logic, exactly one proposition in the set is true; for mathematical simplicity, infinite partitions are ignored) such that $P_{old}(e_i) > 0$ for each i ($1 \leq i \leq n$). Then P_{new} comes from P_{old} by Jeffrey conditionalization with respect to $\{e_1, \ldots, e_n\}$ if and only if every proposition p satisfies:

$$\text{JCOND } P_{new}(p) = \textstyle\sum_{1 \leq i \leq n} P_{new}(e_i)P_{old}(p \mid e_i)$$

Bayesian conditionalization is just the special case where $\{e_1, \ldots, e_n\} = \{e, \sim e\}$ and $P_{new}(e) = 1$.

Jeffrey conditionalization cannot reduce probabilities from 1. If $P_{old}(p) = 1$ then $P_{new}(p) = 1$ by JCOND. The idea is rather that no empirical proposition need acquire probability 1 when one learns from experience. On the approach of this chapter, by contrast, evidence must have evidential probability 1, and some empirical propositions must be evidence if evidential probabilities are ever to change. Should the present approach be modified to permit Jeffrey conditionalization?

The updating of evidential probability by Jeffrey conditionalization is hard to integrate with any adequate epistemology, because we have no substantive answer to the question: what should the new weights $P_{new}(e_i)$ be? Indeed, if sufficiently fine partitions are used, any probability distribution P_{new} is the outcome of any probability distribution P_{old} by JCOND, provided only that $P_{new}(p) = 1$ whenever $P_{old}(p) = 1$ and the set of relevant propositions is finite.[2] Arguably, the same applies to

[2] Proof: Let the propositions of interest be $p_1, \ldots, p_m$. Each of the 2^m possible distributions of truth-values to them corresponds to a conjunction of p_i or $\sim p_i$ for $1 \leq i \leq n$.

BCOND.[3] But there is a simple schematic answer to the epistemological question 'Which instances of BCOND update evidential probability?': those in which e is one's new evidence. Although that answer immediately raises the further question 'What is one's evidence?', it still constitutes progress, for it divides the theoretical labour, allowing other work in epistemology and in philosophy of science—such as Chapter 9—to provide Bayesianism with its theory of evidence. To the parallel question 'Which instances of JCOND update evidential probability?', no such simple answer will do. Jeffrey conditionalization is not conditionalization on evidence-constituting propositions. Moreover, the weights $P(e_i)$ are highly sensitive to background knowledge. When I see a cloth by candlelight, the new probability that it is green depends on my prior knowledge about its colour, the reliability of my eyesight, and the lighting conditions. Attempts to isolate an evidential input in JCOND have not met with success (see Jeffrey 1975, Field 1978, Garber 1980, and Christensen 1992). Jeffrey conditionalization seems not to admit the kind of articulation which would allow work in other areas of epistemology and of philosophy of science to provide it with a standard of appropriateness for the weights. Without such a standard, an account based on Jeffrey conditionalization promises little epistemological insight.

These 2^m conjunctions form a partition. Perhaps $P_{old}(g) = 0$ for some such conjunction g; by disjoining each such g with a conjunction f such that $P_{old}(f) > 0$, form a partition $\{e_1, \ldots, e_n\}$ such that $P_{old}(e_i) > 0$ for $1 \leq i \leq n$. Each e_i is equivalent to a disjunction $f_i \vee g_i$, where $P_{old}(f_i) > 0$, $P_{old}(g_i) = 0$, and f_i either entails p_j or entails $\sim p_j$ $(1 \leq j \leq n)$. Since $P_{old}(g_i) = 0$, standard reasoning shows that $P_{old}(e_i \equiv f_i) = 1$, so $P_{old}(p_j|e_i) = P_{old}(p_j|f_i)$. Thus $P_{old}(p_j|e_i)$ is 1 or 0, depending on whether f_i entails p_j or $\sim p_j$. Now suppose that for every proposition q, if $P_{old}(q) = 1$ then $P_{new}(q) = 1$. Then $P_{new}(g_i) = 0$, and parallel reasoning shows that if $P_{new}(e_i) > 0$ then $P_{new}(p_j|e_i)$ is 1 or 0, depending on whether f_i entails p_j or $\sim p_j$. Thus if $P_{new}(e_i) \neq 0$, $P_{new}(p_j|e_i) = P_{old}(p_j|e_i)$ $(1 \leq i \leq n)$. From this, JCOND is a routine corollary, with any p_j in place of p.

[3] It depends on whether one can introduce finer distinctions than those made by the propositions of interest. If not, and only two possibilities can be distinguished, then no Bayesian conditionalization can change P_{old} to P_{new} where $0 < P_{old}(p) < P_{new}(p) < 1$, because the proposition conditionalized on either makes no difference or eliminates one possibility, in which case all probabilities go to 0 or 1. If finer distinctions can be introduced, $P_{new}(p) = 1$ whenever $P_{old}(p) = 1$, and the set of propositions of interest is finite, then P_{new} comes by BCOND from an extension of P_{old} to the new partition. For suppose that $\{e_1, \ldots, e_n\}$ is a partition such that P_{old} and P_{new} are defined only on propositions equivalent to disjunctions of the e_i. Since $P_{new}(p) = 0$ whenever $P_{old}(p) = 0$, there is a real number $c > 0$ such that for all p for which the probabilities are defined, $cP_{new}(p) \leq P_{old}(p)$. Introduce a new proposition f, bifurcating e_i into $e_i \wedge f$ and $e_i \wedge \sim f$. Determine a probability distribution P_* for the refined partition by: $P_*(e_i \wedge f) = cP_{new}(e_i)$; $P_*(e_i \wedge \sim f) = P_{old}(e_i) - cP_{new}(e_i)$. Hence whenever the probabilities are defined, $P_*(p \wedge f) = cP_{new}(p)$, $P_*(p \wedge \sim f) = P_{old}(p) - cP_{new}(p)$, and $P_*(p) = P_{old}(p)$; thus P_* extends P_{old}. Now $P_*(p|f) = P_*(p \wedge f)/P_*(f) = cP_{new}(p)/c = P_{new}(p)$. Thus P_{new} comes from P_* by BCOND wherever P_{new} is defined. See further Diaconis and Zabell (1982).

Jeffrey evades the normative question by emphasizing the involuntariness of perceptual beliefs. He denies that sense experience provides *reasons* for belief: it is a mere cause, and none the worse for that (Jeffrey 1983): 184–5). However, normative questions arise even for involuntary beliefs. When the sight of a black cat causes a superstitious man to believe that disaster is about to strike, it may be improbable on his evidence that disaster is about to strike. Although most perceptual beliefs are involuntary, Jeffrey himself is willing to judge them by norms, for he regards Bayesianism as a normative theory, not a descriptive one (Jeffrey 1983: 166–7).

Part of the rationale for Jeffrey conditionalization may also depend on an impoverished theory of propositions. Jeffrey's motivating example involves colour vision in poor light; he argues that no proposition 'expressible in the English language' can 'convey the precise quality of the experience' (1983: 165). Surely no context-independent English sentence conveys the precise quality of the experience. It is much less obvious that in the given context no English sentence with perceptual demonstratives (for example, 'It looks like *that*') can express a proposition which would convey the precise quality of the experience, in the sense that Bayesian conditionalization on it would capture the evidential upshot of the experience (see Christensen 1992, but also section 9.5).

The problem about the certainty of evidence arose from the combination of two claims:

> PROPOSITIONALITY The evidential probability of a proposition is its probability conditional on the evidence propositions.
>
> MONOTONICITY Once a proposition has evidential probability 1, it keeps it thereafter.

For PROPOSITIONALITY entails that evidence propositions have evidential probability 1, which by MONOTONICITY implies that they have that status ever after, which is epistemologically implausible. Accounts based on Jeffrey conditionalization retain MONOTONICITY but reject PROPOSITIONALITY; however, they do not yield a non-empty account of evidential probability. A more promising strategy is to retain PROPOSITIONALITY and reject MONOTONICITY. It will be pursued here. PROPOSITIONALITY will henceforth be assumed.

Both BCOND and JCOND allow propositions to acquire probability 1, but not to lose it. They are asymmetric between past and future. Thus a model on which all updating is by Jeffrey or Bayesian conditionalization embodies the empirical assumption that evidence is cumulative, in the sense of MONOTONICITY. In many cases this assumption is false. Bayesians have forgotten forgetting. I toss a coin, see it land heads, put it back in my pocket and fall asleep; once I wake up I have forgotten how it landed. When I saw it land heads, the proposition *e* that it landed heads was part of my evidence; *e* had probability 1 on my evidence. Once I awake, *e* presumably has probability 1/2 on my evidence. No sequence of Bayesian or Jeffrey conditionalizations produced this change in my evidential probabilities. Yet I have not been irrational. I did my best to memorize the result of the toss, and even tried to write it down, but I could not find a pen, and the drowsiness was overwhelming. Forgetting is not irrational; it is just unfortunate. MONOTONICITY is sometimes a useful idealization; it is not inherent in the nature of rationality.

Information loss has a decision-theoretic interest. Before I fall asleep, I am certain that when I wake up I shall have forgotten how the coin landed (I always forget that kind of thing). I am now happy to accept a bet on which I gain £1 if it lands heads and lose £10 otherwise. Tomorrow I shall be happy to accept a bet on which I lose £5 if it lands heads and gain £6 otherwise. If I make both bets, I lose £4 however it lands. I know now that I am vulnerable to such a diachronic Dutch Book, but what can I do? To avoid it by refusing the first bet is just to turn down a certain £1 (compare Skyrms 1993).[4]

A proposition can lose the status of evidence for me even when in the usual sense I forget nothing. Recall an example from section 9.7. I see one red and one black ball put into an otherwise empty bag, and am asked the probability that on the first ten thousand draws with replacement a red ball is drawn each time. I reply '$1/2^{10,000}$'. Part of my evidence is the proposition *e* that a black ball was put into the bag; my calculation relies on it. Now suppose that on the first ten thousand draws a red ball *is* drawn each time, a contingency which my evidence does not rule out in advance, since its evidential probability is non-zero. But when I have seen it happen, I will rationally come to doubt *e*; I will falsely suspect that the ball only looked black by a trick of the light. Thus *e* will no longer form part of my evidence. The traditionalist claim that the possibility of later doubt shows that *e* never was part of my evidence presupposes an untenably Cartesian epistemology.

[4] The case of forgetting shows that not even strong assumptions about the subject's rationality block all counterexamples to the principle of reflection in Van Fraassen 1984. Talbott 1991 has an example of forgetting; the treatment of it by Van Fraassen 1995: 22 cannot plausibly be extended to the present case. For further discussion see Skyrms 1987, Christensen 1991 and 1996, Green and Hitchcock 1994, Howson 1996, Castell 1996, and Hild 1997. Isaac Levi rejects MONOTONICITY in his 1967 and many other publications.

On standard Bayesian accounts of updating, the only present trace of past evidence is in present probabilities. No separate record is kept of evidence, off which a proposition can be struck. But a theory of evidential probability can keep separate track of evidence and still preserve much of the Bayesian framework.[5] Let P be the prior probability distribution, e_w the conjunction of all old and new evidence for one in a case α, and $P_\alpha(p)$ the evidential probability of a proposition p for one in α. The proposal is that P_α is the conditionalization of P on e_α:

$$\text{ECOND} \quad P_\alpha(p) = P(p \mid e_\alpha) = P(p \wedge e_\alpha)/P(e_\alpha) \qquad (P(e_\alpha) > 0)$$

ECOND formalizes PROPOSITIONALITY. It allows MONOTONICITY to fail, for if one forgets something between t and a later time t^*, being in cases α and α^* at t and t^* respectively, then e_{α^*} need not entail e_α, so possibly $P_{\alpha^*}(e_\alpha) < 1$ even though $P_\alpha(e_\alpha) = 1$. Thus a proposition can decrease in probability from 1. In that sense, evidence need not be certain.

When no evidence is lost between α and α^*, e_{α^*} is equivalent to $e_\alpha \wedge f$, where f is the conjunction of the new evidence gained in that interval, and ECOND implies that P_{α^*} results from conditionalizing P_α on the new evidence f. Formally, for any proposition p:

$$\begin{aligned} P_{\alpha^*}(p) &= P(p \wedge e_\alpha \wedge f)/P(e_\alpha \wedge f) = (P(p \wedge e_\alpha \wedge f)/P(e_\alpha))/(P(e_\alpha \wedge f)/P(e_\alpha)) \\ &= P_\alpha(p \wedge f)/P_\alpha(f) = P_\alpha(p \mid f) \end{aligned}$$

BCOND is the special case of ECOND when evidence is cumulative. Thus Bayesian conditionalization can be recovered when needed.

The distribution P is conceptually rather than temporally prior; it need not coincide with P_α for any case α in which some subject is at some time, for P is not a distribution of credences, and the subject may have non-trivial evidence at every time. An incidental advantage of this approach is that it helps with the problem of *old evidence* (Glymour 1980: 85–93, Earman 1992: 119–35, Howson and Urbach 1993: 403–8, and Maher 1996). One would like to say that e confirms h if and only if the conditional probability of h on e is higher than the unconditional

[5] Compare the notion of a diary in Skyrms 1983. Skyrms's discussion concentrates on the problem of memory storage, but remembering which propositions are evidence is no worse than remembering a probability for each proposition. Of course, it is often rational to retain a belief even when one has forgotten one's past evidence for it. In some cases the belief itself has attained the status of evidence (see section 10.3); in others one has only indirect evidence for it (for example, one seems to remember p and is usually right about such things). But even those beliefs are evidentially probable at t only if one's evidence at t supports them. See Harman 1986 for much relevant discussion of clutter avoidance.

probability of h (compare EV in section 9.2). If e is already part of the evidence then its probability is 1, and the conditional probabilities are identical; yet old evidence does sometimes confirm hypotheses. Appeals are sometimes made to probabilities in past or counterfactual circumstances in which the evidence does not include e, but they produce anomalous results, because the evidence in those circumstances may be distorted by irrelevant factors.

Example: a coin is tossed ten times. Let h be the hypothesis that it landed the same way each time. The initial probability of h is $1/2^9$. Witness A says 'I saw the first six tosses; it landed heads each time'. Witness B then says 'I saw the last four tosses; it landed tails each time'; let e be the proposition that B said this. We have no reason to doubt A and B; if they are both telling the truth, then h is false. But B's statement causes A to break down; he admits that he was lying, and has no relevant knowledge. If B had not made his statement, A would not have withdrawn his, and there would have been no reason to suspect that he was lying. Thus, in the nearest past or counterfactual circumstances in which e was not part of our evidence, the conditional evidential probability of h on e is lower than the unconditional evidential probability of h. Nevertheless, in our present situation, e does confirm h, for since we still have no reason to doubt B, the probability of h on our evidence is around $1/2^6$. Once we have the prior probability distribution P, we can say that $P(h \mid e) > P(h)$. If we like, we can relativize confirmation to background information f by requiring that $P(h \mid e \wedge f) > P(h \mid f)$, but this does not justify subjecting it to the vagaries of the evidence we once or would have had. Of course, these remarks are schematic, but at least the general form of the solution does not introduce the irrelevant complications consequent on an identification of the probabilities with past or counterfactual credences.

10.3 EVIDENCE AND KNOWLEDGE

Which propositions are one's evidence? Without a substantive conception of evidence, probabilistic epistemology is empty; in practice, it has taken the existence of such a conception for granted without itself supplying one.

Different conceptions of evidence are compatible with ECOND. Chapter 9 defended the simple, natural proposal that one's evidence is one's body of knowledge. More precisely, one's total evidence e_α in a case α is the conjunction of all the propositions which one knows in α

(E = K).[6] Here 'one' may refer to an individual or a community. Since evidence can lose probability 1, the defeasibility of knowledge by later evidence is no objection to E = K. When I see the black ball put into the bag, the proposition that a black ball was put into the bag becomes part of my evidence because I know that a black ball was put into the bag. When I have seen a red ball drawn each time on the first ten thousand draws, that further evidence undermines my knowledge that a black ball was put into the bag, and the previously known proposition ceases to be part of my evidence. Since only true propositions are known, evidence consists entirely of true propositions, but one true proposition can cast doubt on another.

Subjective Bayesians might identify one's evidence with one's beliefs (understood as propositions of subjective probability 1) rather than with one's knowledge (E = B). Given E = B, one can manufacture evidence for one's favourite theories by manipulating oneself into a state of certainty about appropriate propositions—for example, that one has just seen one's guru perform a miracle. That does not capture the spirit of the injunction to proportion one's belief to one's evidence.

The positive argument for E = K will not be rehearsed here. The rest of the chapter develops the conjunction of E = K with ECOND as a theory of evidential probabilities, in a way which indicates at least their mutual coherence. The concept *knowledge* is sometimes regarded as a kind of survival from stone-age thinking, to be replaced by probabilistic concepts for the purposes of serious twentieth-century epistemology. That view assumes that the probabilistic concepts do not depend on the concept *knowledge*. If E = K and ECOND are true, that assumption is false. The concepts *knowledge* and *evidential probability* are complementary; neither can replace the other.

Some initially surprising results of the theory stem from the point that we are not always in a position to know whether we know something. By E = K, we are not always in a position to know whether something is part of our evidence. Let us briefly rehearse the context in which this consequence is independently plausible. Whether something is part of our evidence does not depend solely on whether we believe it to be part of our evidence. That *p* is part of our evidence is a non-trivial condition; arguably, no non-trivial condition is such that whenever it obtains one is in a position to know that it obtains (see Chapters 4 and 8). But if we are not always in a position to know whether something is part of our evidence, how can we use evidence? We shall sometimes not

[6] Restrictive views of evidence can make unnecessary problems for conditionalization by not allowing propositions about the subject's updated belief state to count as part of the new evidence; this may explain the cases discussed in Howson 1996 and Castell 1996.

be in a position to know the probability of a proposition on our evidence. How then can we follow the rule 'Proportion your belief in a proposition to its probability on your evidence'?

As noted in earlier chapters, there is a recurrent temptation to suppose that we can follow a rule only if it is always cognitively transparent to us whether we are complying with it. On this view, if we are sometimes not in a position to know whether we are ϕ-ing when C, then we cannot follow the rule 'ϕ when C'; at best we can follow the rule 'Do what appears to you to be ϕ-ing when it appears to you that C'. For instance, we cannot follow the rule 'Add salt when the water boils' because we are not always in a position to know whether something is really salt, water, or boiling; at best we can follow the rule 'Do what appears to you to be adding salt when what appears to you to be water appears to you to boil'. Can we even follow the modified rule? That something appears to us to be so is itself a non-trivial condition. But we *can* follow the rule 'Add salt when the water boils', even though we occasionally make mistakes in doing so. It is enough that we *often* know whether the condition obtains. Compliance with a non-trivial rule is never a perfectly transparent condition. We use rules about evidence for our beliefs because they are often less opaque than rules about the truth of our beliefs; perfect transparency is neither possible nor necessary.

Just as we can follow the rule 'Add salt when the water boils', so we can follow the rule 'Proportion your belief in a proposition to its probability on your evidence'. Although we are sometimes reasonably mistaken or uncertain as to what our evidence is and how probable a proposition is on it, we often enough know enough about both to be able to follow the rule. It is easier to follow than 'Believe a proposition if it is true', but not perfectly easy. And just as adding salt when the water boils is not equivalent to doing one's rational utmost to add salt when the water boils, so proportioning one's belief in a proposition to its probability on one's evidence is not equivalent to doing one's rational utmost to proportion one's belief in a proposition to its probability on one's evidence. The content of a rule cannot be reduced to what it is rational to do in attempting to comply with it. Evidential probabilities are not rational credences.

The next task is to develop a formal framework for the combination of E = K with ECOND, by appropriating some ideas from epistemic logic.[7] Within this framework, the failure of cognitive transparency for evidential probabilities will receive a formal analysis.

[7] The application of modal logical techniques to epistemological problems was pioneered in Hintikka 1962, although the assumptions made here differ from Hintikka's. A good text for the modal logical background is Hughes and Cresswell 1996.

10.4 EPISTEMIC ACCESSIBILITY

For the sake of familiarity, we may speak of notional *worlds* rather than cases. In order to facilitate discussion of intersubjective knowledge, we do not conceive a world as centred on a subject and a time. Rather, we implicitly specify the epistemic perspective by our choice of an accessibility relation between worlds (see below). We assume a set of mutually exclusive and jointly exhaustive worlds. In a given application, worlds need be specific only in relevant respects. We need not assume that all worlds are metaphysically possible, in the sense that they could really have obtained. A set of all worlds is assumed. The relevant propositions are true or false in each world, and closed under truth-functional combinations. We assume that for each set of worlds, some proposition is true in every world in the set and false in every other world.

Let P be a prior probability distribution as in section 10.1. P is assumed to satisfy the axioms of the probability calculus as stated in terms of worlds. Thus $P(p) = 1$ whenever p is true in every world; $P(p \vee q) = P(p) + P(q)$ whenever p and q are in no world both true. Consequently, if p and q are true in exactly the same worlds, $P(p) = P(q)$.[8] For any set of worlds, some proposition is true at exactly the worlds in the set, and all such propositions are equiprobable; thus the assignment of probabilities to propositions induces a unique assignment of probabilities to set of worlds. Conversely, an assignment of probabilities to sets of worlds induces a unique assignment of probabilities to propositions.

Propositions are known or not known in worlds; propositions about which propositions one knows are true or false in worlds. The account will not assume any general principle about knowledge, except that a proposition is true in any world in which it is known. In particular, it will not assume logical omniscience; if p and q are true in exactly the same worlds, one may know p and not know q. Relative to a subject S and a time t, a world x is *epistemically accessible* ('accessible' for short) from a world w if and only if every proposition which S knows at t in w is true in x. A world is accessible if, for all one knows, one is in it. Since knowledge implies truth, every world is accessible from itself. A proposition p is consistent with propositions $q_1, \ldots, q_n$ if and only if all of p and $q_1, \ldots, q_n$ are true in some world; thus, in a world w, p is *consistent with what one knows* if and only if p is true in some world accessible

[8] If p and q are true in exactly the same worlds, then in no worlds are both p and $\sim q$ true, and $p \vee \sim q$ is true in all worlds, so by the axioms $1 = P(p \vee \sim q) = P(p) + P(\sim q)$. By the same reasoning, $1 = P(q) + P(\sim q)$. Thus $P(p) = 1 - P(\sim q) = P(q)$.

from w (compare the standard possible worlds semantics for the possibility operator ◊). Similarly, p follows from $q_1, \ldots, q_n$ if and only if p is true in every world in which all of $q_1, \ldots, q_n$ are true; thus, in a world w, *p follows from what one knows* if and only if p is true in every world accessible from w (compare the standard possible worlds semantics for the necessity operator □). Trivially, if one knows a proposition then it follows from what one knows, but the converse may fail, since one need not know that which follows from what one knows.

Now assume ECOND and E = K; in all worlds, evidential probabilities are probabilities conditional on one's evidence and one's evidence is what one knows. Relative to a subject S at a time t, for any world w, e_w is the conjunction of S's evidence at t in w. By E = K, e_w is true in all and only the worlds accessible from w. P_w is the distribution of evidential probabilities for one in w. ECOND says that P_w results from conditionalizing the appropriate prior distribution P on e_w.

When the set of worlds is at most countably infinite, a further natural constraint on P is that it be *regular*, in the sense that $P(p) = 0$ only if p is true in no world: the probability distribution does not rule out any world in advance. When there are uncountably many worlds, no probability distribution is regular (infinitesimal probabilities are not being considered here). The most natural prior distributions are those for which there is a finite number n of worlds, and $P(p) = m/n$ whenever p is true in exactly m worlds, but such uniformity in P will not be assumed. Since knowledge entails truth, e_w is always true in w. Thus when P is regular, $P(e_w) > 0$ for each w, so probabilities conditional on e_w are well defined and ECOND defines evidential probabilities everywhere. Regularity also entails that the evidential probability of p is 1 only if p follows from one's evidence, for if p is false in some world in which e_w is true, then $P(\sim p \wedge e_w) > 0$, so $P_w(p) < 1$. Regularity likewise entails that p follows from what one knows if and only if the evidential probability of p is 1, and that p is consistent with what one knows if and only if the evidential probability of p is non-zero.

Propositions about evidential probability are themselves true or false in worlds. For example, the proposition that p is more probable than not on the evidence is true in w if and only if $P_w(p) > 1/2$. Thus propositions about evidential probability themselves have probabilities.[9]

In the manner of possible worlds semantics, conditions on accessibility

[9] See Skyrms 1980 and Gaifman 1988 for interesting discussions of higher-order probability. Their subjectivism introduces complications into their accounts. For example, Gaifman needs a distinction between the agent's probability and a hypothetical expert's probability to handle higher-order probability. These complications are unnecessary from the present perspective.

correspond to conditions on knowledge, which in turn have implications for evidential probabilities. For example, accessibility is transitive if and only if for every proposition *p* in every world, if *p* follows from what one knows then that *p* follows from what one knows itself follows from what one knows (compare the S4 axiom $\Box p \supset \Box\Box p$ in modal logic). The latter condition follows from the notorious 'KK' principle that when one knows *p*, one knows that one knows *p*; it is slightly weaker, but not weak enough to be true, even for all rational subjects (see Chapter 5). For a regular probability distribution, transitivity is equivalent to the condition that when *p* has evidential probability 1, the proposition that *p* has evidential probability 1 itself has evidential probability 1.

Accessibility is symmetric if and only if for every proposition *p* in every world, if *p* is true then that *p* is consistent with what one knows follows from what one knows (compare the Brouwersche axiom $p \supset \Box\Diamond p$). For a regular probability distribution, symmetry is equivalent to the condition that when *p* is true, the proposition that *p* has non-zero evidential probability itself has evidential probability 1. There is good reason to doubt that accessibility is symmetric. Let *x* be a world in which one has ordinary perceptual knowledge that the ball taken from the bag is black. In some world *w*, the ball taken from the bag is red, but freak lighting conditions cause it to look black, and everything which one knows is consistent with the hypothesis that one is in *x*. Thus *x* is accessible from *w*, because every proposition which one knows in *w* is true in *x*; but *w* is not accessible from *x*, because the proposition that the ball taken from the bag is black, which one knows in *x*, is false in *w*. Let *p* be the proposition that the ball taken from the bag is red. In *w*, *p* is true, but that *p* is consistent with what one knows does not follow from what one knows, for what one knows is consistent with the hypothesis that one knows ~*p* (see section 8.2 and Humberstone 1988 for related issues). On a regular probability distribution, the evidential probability in *w* of the proposition that *p* has non-zero evidential probability falls short of 1 in this case.

Such examples depend on less than Cartesian standards for knowledge and evidence; Bayesian epistemology must learn to live with such standards. Moreover, failures of symmetry can result from processing constraints, even when false beliefs are not at issue (see also Shin and Williamson 1994). For a crude example, imagine a creature which knows all the propositions recorded in its memory; we may pretend for simplicity that it is somehow physically impossible for false propositions to be recorded there. Unfortunately, there is no limit to the time taken to deliver propositions from memory to the creature's central processing unit. Now toadstools are in fact poisonous for the creature, but it has no memory of any proposition relevant to this truth. It wonders whether it knows that toadstools are not poisonous. It searches for relevant memories. At any time, it has recovered no relevant memory, but for all it knows that is merely because the delivery procedure is slow, and in a moment the memory that toadstools are not poisonous will be delivered, in which case it will have known all along that they are not poisonous. Everything which it knows in the actual world *w* is true in a world *x* in which it knows that toadstools are not poisonous; thus *x* is accessible from *w*. But *w* is not accessible from *x*, because something which it knows in *x* (that toadstools are not poisonous) is false in *w*. Although in *w* the proposition *p* that toadstools are poisonous is true, that *p* is consistent with what it knows does not itself follow from what it knows.

Epistemic logic and probability theory are happily married because the posterior probabilities in *w* result from conditionalizing on the set of worlds epistemically accessible from *w*. This idea has become familiar in standard applications of epistemic logic to the concept of *common knowledge* in decision theory and game theory (see for example Fudenberg and Tirole 1991: 541–72). As usual, the proposition that *p* is common knowledge is analysed as the infinite conjunction of *p*, the proposition that everyone knows *p*, the proposition that everyone knows that everyone knows *p*, and so on. Thus the analysis of common knowledge requires an account of knowledge. Something like the framework above is used, with a separate accessibility relation R_S for each agent S but a common prior probability distribution; different agents can have different posterior probabilities in the same world because they have different sets of accessible worlds on which to conditionalize. 'S knows *p*' (K_Sp) is given the semantics of '*p* follows from what one knows' with respect to the accessibility relation R_S; thus knowledge is treated as closed under logical consequence (contrast the present account). Furthermore, in decision theory accessibility is usually required to be an equivalence relation (symmetric and transitive as well as reflexive) for each agent. On this model, the agent partitions the set of worlds into a set of mutually exclusive and jointly exhaustive sets. In *w*, the agent knows just those propositions which are true in every world belonging to the same member of the partition as *w*. Informally, imagine that each world presents a particular appearance to the agent, who knows all about appearances and nothing more; thus one world is epistemically accessible from another if and only if they have exactly the same appearance, which is an equivalence relation. The corresponding propositional logic of knowledge is the modal system S5, with K_S in place of $\Box$; one can axiomatize it by taking as axioms all truth-functional tautologies and formulas of the forms $K_S(A \supset B) \supset (K_SA \supset K_SB)$, $K_SA \supset A$, and $\sim K_SA \supset K_S\sim K_SA$, and as rules of inference modus ponens and epistemization (if A is a theorem, so is K_SA). One of the earliest results to be proved on the basis of assumptions tantamount to these was Aumann's '[no] agreeing to disagree' theorem: when the posterior probabilities of *p* for two agents are common knowledge, they are identical (Aumann 1976; the proof relies heavily on the assumption of common prior probabilities).

Earlier examples expose some of the idealizations implicit in the partitional model of knowledge. In particular, the counterexamples to the symmetry of accessibility, and so to the Brouwersche schema $\sim A \supset K_S\sim K_SA$, are equally counterexamples to the S5 schema $\sim K_SA \supset K_S\sim K_SA$, given the uncontentious principle that knowledge implies truth ($K_SA \supset A$). Some progress has been made in generalizing results such as Aumann's to weaker assumptions about knowledge (Bacharach 1985, Geanakoplos 1989, 1992, 1994, Samet 1990, Shin 1993, Basu 1996). It can be argued that, even when logical omniscience is assumed, the propositional logic of knowledge is not S5 but the modal system KT (alias T), which one can axiomatize by dropping the axiom schema $\sim K_SA \supset K_S\sim K_SA$ from the axiomatization above (Williamson 1994b: 270–5). What KT assumes about knowledge, in addition to logical omniscience, is just that knowledge implies truth. When K_SA is read as 'It follows from what one knows that A', rather than as 'One knows that A' (where one is S), logical omniscience becomes unproblematic for K_S, whatever S's logical imperfections.

10.5 A SIMPLE MODEL

We can gain a more intuitive feel for the present account of higher-order probabilities by working through some of its consequences in a toy example. In doing so we can combine it with the account of margins for error in Chapter 5.

According to a straightforward margin for error principle, S knows *p* in a world *w* only if *p* is true in every world sufficiently close to *w* in the relevant respects (which will depend on the particular case). In the simplest models, that condition is sufficient as well as necessary for knowing *p*: K*p* is true in *w* if and only if *p* is true in all worlds close to *w* (for simplicity, we omit the subscript 'S'). Let us introduce an operator B, where B*p* is to mean that *p* is highly probable on S's evidence. On the present account of evidence, we can then say that in such a model, Bp is true in w if and only if p is true in *most* worlds close to w. That is only a first approximation, of course, because some worlds close to w may be assigned higher probabilities than others, and 'most' is problematic for infinite sets; but we can build the required probability distribution into our understanding of 'most'. The result is a *probabilistic margin for error principle*. Whereas any operator defined by the original margin for error principle is automatically factive, because every world is close to itself, B is not in general factive, because a set which contains most worlds close to w need not contain w itself. A false proposition can be highly probable on one's evidence; some evidence is misleading. In place of the factiveness principle K$p \supset p$, one can expect only the weaker consistency principle B$p \supset$ ~B~p; two disjoint sets cannot each contain most worlds close to w. Contradictories cannot both be highly probable on one's evidence.

For definiteness, we can imagine the worlds of our toy model as forming a two-dimensional infinite grid. For convenience, each world may be identified with a 'point', a pair of coordinates $<x,y>$, where x and y are any integers. Again for convenience, we may identify propositions with sets of points; a proposition is true at a point if and only if the latter belongs to the former. Let us count the points close to $<x,y>$, the points accessible from it, as just those within one step of it on the grid: $<x,y>$ itself, $<x+1,y>$, $<x-1,y>$, $<x,y+1>$ and $<x,y-1>$. All worlds are treated as equiprobable. Let us count most of these five points as in a set if and only if at least four are. Thus the proposition Bp is true in a world $<x,y>$ if and only if $\{<u,v>: |x-u|+|y-v| \leq 1\} \cap p$ has at least four members, whereas Kp is true in $<x,y>$ if and only if $\{<u,v>: |x-u|+|y-v| \leq 1\} \cap p$ has five members. We can read B as 'It is at least 80 per cent probable that', understanding 'probable' evidentially.

As an example, let p be the proposition $\{<0,1>, <1,0>, <1,2>, <2,1>\}$. The only point close to at least four members of p is $<1,1>$, so Bp is $\{<1,1>\}$. No point is close to at least four members of Bp, so BBp is $\{\}$. Thus $<1,1>$ is a point at which p is false but at least 80 per cent probable, although it is only 20 per cent probable that p is at least 80 per cent probable. This illustrates the simultaneous breakdown of factiveness and the BB principle that if p is at least 80 per cent probable then it is at least 80 per cent probable that p is at least 80 per cent probable. More generally in the model, B is subject to erosion effects typical of margin for error principles. For example, if p is any finite set, then $\mathrm{B}^k p$ (k iterations of B on p) is empty for some natural number k.[10]

[10] Proof: B$p = \{<x,y>: \{<u,v>: |x-u|+|y-v| \leq 1\} \cap p$ has at least 4 members$\}$. If p is finite then $p \subseteq \{<x,y>: |x-a|+|y-b| < k\}$ for some pair $<a,b>$ and natural number k. Then if

As required, Bp entails ~B~p in the model. Also as we should expect, the closure principle that if $p_1, \ldots, p_n$ logically entail q then B$p_1, \ldots,$ Bp_n logically entail Bq holds when $n \leq 1$ but not otherwise. For example, if p_1 is {<0,1>, <1,0>, <1,2>, <2,1>} and p_2 is {<1,0>, <1,1>, <1,2>, <2,1>}, then both Bp_1 and Bp_2 are {<1,1>}, but $p_1 \wedge p_2$ is the intersection of p_1 and p_2, {<1,0>, <1,2>, <2,1>}; since this has only three members, B($p_1 \wedge p_2$) is {}. Each of p_1 and p_2 is 80 per cent probable at <1,1>, but their conjunction is only 60 per cent probable. By contrast, K satisfies the corresponding closure principle in this model for multi-premise inference.

Another contrast between strict and probabilistic margins for error in this model is that K satisfies the Brouwersche principle $p \supset$ K~K~p because closeness is symmetric, but B does not satisfy the corresponding principle $p \supset$ B~B~p. For example, if p is {<1,1>}, then ~B~p is {}, so B~B~p is {}.

The exposition of the present theory of probabilities on evidence is now complete, and some readers may wish to skip the rest of this chapter. However, deviations from the partitional model sketched at the end of section 10.4 generate a phenomenon which seems to threaten the proposed marriage of knowledge and probability. The aim of the next section is to understand that phenomenon.

10.6 A PUZZLING PHENOMENON

The paradoxical phenomenon can be illustrated thus. There are just three worlds: w_1, w_2 and x. As in Figure 3, x is accessible from each world; each of w_1 and w_2 is accessible only from itself. Thus accessibility is reflexive and transitive, but not symmetric. For simplicity, the subject will be treated as logically omniscient; the paradoxical phenomenon does not depend on the failure of knowledge to be deductively closed. Since the only world accessible from x is x itself, if one is in x then one knows that one is in x. Since the worlds accessible from w_i are x and w_i, if one is in w_i then one knows that one is in either w_i or x, but one does not know which; for all one knows, one *knows* that one is in x. In w_i, although one is not in x, and therefore does not know that one is in x, one does not know that one does not know that one is in x. This is just

$k-1 \leq |x-a|+|y-b|$, $<x,y> \notin$ Bp: for example, if $a \leq x$ and $b \leq y$, then $k \leq |x+1-a|+|y-b|$ and $k \leq |x-a|+|y+1-b|$, so $<x+1,y> \notin p$ and $<x,y+1> \notin p$, so $\{<u,v>: |x-u|+|y-v| \leq 1\} \cap p$ has at most 3 members. Thus B$p \subseteq \{<x,y>: |x-a|+|y-b| < k-1\}$. By induction, B$^k p$ = {}.

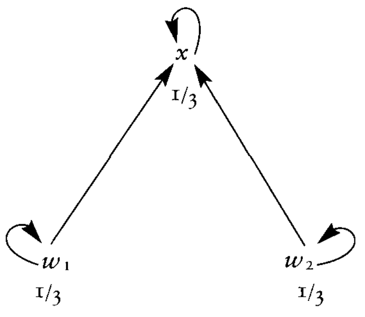

FIGURE 3

the failure of the Brouwersche and S5 axioms for knowledge in a non-symmetric model.

The prior probability distribution is uniform; each world has a prior probability of 1/3. Let *p* be the proposition that one is in w_1 or w_2. The prior probability of *p* is 2/3. If one is in *x*, then *p* is false in all accessible worlds, so its posterior probability is 0. If one is in w_i, then *p* is true in just one of the two accessible worlds, so its posterior probability is 1/2. Thus one knows in advance that the posterior probability of *p* will be either 0 or 1/2, and so in any case lower than its initial probability.[11] But if one knows in advance that, when the evidence comes in, the probability of *p* on the evidence will drop from 2/3 to at most 1/2, why is that known feature of the future evidence not anticipated by lowering the *prior* probability of *p* to at most 1/2? Surely the posterior probabilities are a better guide to the truth than the prior probabilities are, because they are based on more evidence (compare Shin 1989 and 1992 and Geanakoplos 1989, 1992, and 1994).

A money pump argument makes the problem vivid. Consider a ticket which entitles one to £6 if *p* is true and to nothing if *p* is false. The initial probability that the ticket entitles one to £6 is 2/3. Given standard Bayesian decision theory, one should be willing to pay up to 2/3£6 + 1/3£0 = £4 in advance for the ticket. But the posterior probability that the ticket entitles one to £6 is at most 1/2, so once the evidence is in one should

[11] 'Know in advance' here could just mean 'know a priori'. As explained in section 10.2, the prior probabilities need not be the evidential probabilities at some earlier time t_0. But the example is more vivid if the initial probabilities are one's evidential probabilities at t_0. On this reading, the diagram does not illustrate one's world at t_0; it is confined to one's worlds at the relevant later time. The problem arises even if the 'prior' probabilities are not the absolutely prior probabilities in ECOND, for the updating can be regarded as an instance of BCOND, which is formally a special case of ECOND.

be willing to sell the ticket for any price from 1/2£6 + 1/2£0 = £3 upwards. Indeed, if the evidence shows that one is in x, then one knows that the ticket is worthless. A shark can apparently pump money out of one by selling one many such tickets for £4 before the evidence is in and buying them back afterwards for £3. I remain a money pump even if I require a small profit on each transaction. Moreover, one knows all that in advance. Is there not something irrational in such an assignment of probabilities?

Reasons emerged in section 10.1 to deny that decision-theoretic arguments have a direct bearing on evidential probabilities. Such arguments are especially dubious when (as above) the probabilities do not all belong to the same time (see, for example, Christensen 1991). Nevertheless, the money pump argument provides an intuitive framework for generalizing the problem. For simplicity, let the worlds form a finite set W. It will be convenient to treat the bearers of probability as subsets of W. For $w \in W$, let $R(w)$ be the set of worlds to which w bears the accessibility relation R. Since e_w is true in exactly the worlds in $R(w)$, the posterior probability $P_w(X)$ of X in w is $P(X|R(w))$ by ECOND ($X \subseteq W$). The expectation $E(P_w(X))$ of the evidential probability random variable $P_w(X)$ is therefore $\Sigma_{w \in W}\ P(\{w\})P(X|R(w))$. The identity of prior and expected posterior probabilities comes to this:

$$\text{EXP}\quad P(X) = \sum_{w \in W} P(\{w\})P(X|R(w))$$

Consider a ticket which entitles one to £n if one's world is in X and to nothing otherwise. Suppose that before the evidence is in one buys the ticket at its expected (monetary) value at that time; after the evidence is in one sells the ticket at its expected value at that later time. What is one's expected profit or loss over the two transactions? The buying price is P(X)£n. The expected selling price is the expected posterior probability of X times £n. In the example above, the prior probability of p was 2/3; its expected posterior probability was 1/3 (0) + 2/3(1/2) = 1/3; the expected profit was 1/3 £6−2/3£6, a loss of £2. Thus, if the left-hand side of EXP is less or greater than its right-hand side, one's expected profit over the two transations is positive or negative respectively. Since the prior and expected posterior probabilities of W−X are one minus the prior and expected posterior probabilities of X respectively, an expected profit on the two transactions with respect to X implies an expected loss on the corresponding transactions with respect to W−X. Thus unless EXP holds, the transactions make one a money pump with respect to some proposition (see Goldstein 1983, Van Fraassen 1984, and Skyrms 1987 for related discussion).

One response to the strange situation is to deny that it can arise. On this view, the money pump argument shows that no probability distribution P on a set of worlds W with an epistemic accessibility relation R can violate EXP for any $X \subseteq W$; Figure 3 does not picture a genuine possibility. It can be proved that, given a relation R on a finite set W, EXP holds for every regular probability distribution P on W and $X \subseteq W$ if and only if R is an equivalence relation on W (Appendix 4, proposition 5). In partitional models of knowledge, expected posterior probabilities always coincide with prior probabilities; any deviation from partitionality makes them diverge on a suitable probability distribution.[12]

Although the interpretation of R as an accessibility relation for knowledge automatically requires R to be reflexive, one cannot escape the result just by allowing R to be non-reflexive and reinterpreting it as an accessibility relation for (say) rational belief, in the sense that *x* is accessible from *w* if and only if whatever the subject rationally believes in *w* is true in *x*. The aforementioned result holds provided that each member of W has R to at least one member of W, not necessarily itself: in other words, provided that R is serial. The accessibility relation for rational belief is non-serial only when (if ever) rational beliefs are inconsistent. In that case $R(w)$ is sometimes empty, so the expected posterior probability is not well defined. If R is serial, $R(w)$ is always non-empty; given regularity, the expected posterior probability is then well defined. If the accessibility relation is serial but not reflexive, then expected posterior probabilities diverge from prior probabilities on a suitable probability distribution.

If epistemic accessibility had to be an equivalence relation, EXP would always hold. But the counterexamples to partitionality have not lost their force. Of course, realistic examples involve far more complex epistemic situations than that illustrated above. Nevertheless, we can begin to understand the mechanics underlying non-partitionality by filling out the example above of the worlds w_1, w_2, and x in some detail.

A simple creature monitors the ambient temperature by means of two detectors. When it is not cold, the first detector is activated and causes the information that it is not cold to be stored; otherwise the first detector is inactive. When it is not hot, the second detector is activated and causes the information that it is not hot to be stored; otherwise the second detector is inactive. The relevant three (partial) worlds are w_1 (it is

[12] Partitionality is also equivalent to a form of the Principal Principle, or Miller's Principle: $P(X|\{w \in W{:}P(X|R(w)) = c\}) = c$ for every real number c, $X \subseteq W$, and regular probability distribution P on W such that the conditional probability is defined (Appendix 4, proposition 6). Skyrms 1980 explores the relation of such principles to probability kinematics.

hot), w_2 (it is cold), and x (it is neither hot nor cold). In w_1, only the information that it is not cold is stored. In w_2, only the information that it is not hot is stored. In x, both the information that it is not hot and the information that it is not cold is stored. Unfortunately, the creature has no capacity to survey what it has stored and detect that a particular piece of information is not stored. Hence, in w_1 it cannot detect that the information that it is not hot is not stored, and infer that it is hot. Similarly, in w_2 it cannot infer that it is cold. Since it never stores false information, we can reasonably treat it as knowing the stored information and no more. Thus the worlds epistemically accessible from w_1 are w_1 and x; the worlds accessible from w_2 are w_2 and x; the only world accessible from x is x itself. Let the three worlds be equiprobable in advance, and treated as such by the creature. Then the epistemic situation is exactly that depicted in Figure 3. If we like, we can elaborate the story to endow the creature with significant powers of logic and self-reflection (see Appendix 5 for details).

Is the initial assignment of equal probabilities to the three worlds irrational? Would some other initial assignment do better? Let P be a regular prior probability distribution which coincides with the corresponding distribution of expected posterior probabilities. So, in particular:

$$\mathrm{P}(\{x\}) = \Sigma_{y\in \mathrm{W}}\, \mathrm{P}(\{y\})\mathrm{P}(\{x\}|\mathrm{R}(y))$$

By the diagram $x \in \mathrm{R}(y)$ for all $y \in \mathrm{W}$, so $\mathrm{P}(\{x\}|\mathrm{R}(y)) = \mathrm{P}(\{x\})/\mathrm{P}(\mathrm{R}(y))$. Dividing through by $\mathrm{P}(\{x\})$ gives:

$$1 = \Sigma_{y\in \mathrm{W}}\, \mathrm{P}(\{y\})/\mathrm{P}(\mathrm{R}(y))$$

But $\mathrm{R}(x) = \{x\}$, so $\mathrm{P}(\{x\}/\mathrm{P}(\mathrm{R}(x)) = 1$, so:

$$0 = \mathrm{P}(\{w_1\})/\mathrm{P}(\mathrm{R}(w_1)) + \mathrm{P}(\{w_2\})/\mathrm{P}(\mathrm{R}(w_2))$$

Thus $\mathrm{P}(\{w_1\}) = \mathrm{P}(\{w_2\}) = 0$. This contradicts the assumed regularity of P. Only an irregular prior distribution on W can coincide with the corresponding expected posterior distribution. Specifically, the proof shows that either $\mathrm{P}(\{x\}) = 0$, in which case $\mathrm{P}(\{x\}|\mathrm{R}(x))$ is undefined, or $\mathrm{P}(\{w_1\}) = \mathrm{P}(\{w_2\}) = 0$. The creature can align its prior probabilities with its expected posterior probabilities only by ruling out some of the three worlds in advance. But that would be quite irrational; each of them is an epistemically live possibility. The uniform prior distribution was not to blame. One must learn to live with the divergence between prior and expected posterior probabilities: but how?

Consider the money pump argument first. As given above, it assumes that, once the evidence is in, the agent can calculate the relevant expectations, which requires it to know the posterior probabilities. That is just what the structure of the accessibility relation precludes. In the three-world model, it is certain in advance that the posterior probability that it is hot is 1/2 when it is hot and 0 otherwise. Hence if, when it was hot, the creature knew that the posterior probability that it was hot was 1/2, it could deduce that it was hot; but then the posterior probability that it was hot would be 1, not 1/2. For simplicity, sentences about probabilities and actions were omitted from the creature's language; their addition would complicate but not undermine the argument, provided that the creature has no more empirical evidence than before. Thus, when it is hot, the creature cannot know the probability on its evidence that it is hot. It does not know the premises of the decision-theoretic calculation. Even so, the probabilities on its evidence can still play a causal role in its decision-making, for its evidence is physically realized as its stored information. Thus decisions can be made when it is hot that would not have been made if it had not been hot.

Could the creature discover that it is hot by observing its own actions? Once it has acted, it is in a different world; its action may even have changed the temperature. Perhaps it can work out that it *was* hot, but that would not imply that it could have had the present tense knowledge before it acted. Could it have introspected its intention to act in a certain way before carrying it out? Sometimes we do not know whether we are going to act in a certain way until we carry out the action; let the creature be like that when it does not know the probabilities on which it will act.

If we assume that prior probabilities should align themselves with expected probabilities posterior to the future acquisition of knowledge, we assign the probability of being known in the future a privileged status in the present. Why should I give the property of being known by me tomorrow a privileged status today? There is one reason: whatever I shall know tomorrow is true. Thus if I know today that tomorrow I shall know *p*, I can deduce *p* today. By contrast, if I rationally believe today that tomorrow I shall rationally believe *p*, I cannot deduce *p* today; for all I rationally believe today, tomorrow's rational belief will be based on misleading evidence.[13] But this is no reason to give the property of being known by me tomorrow a more privileged status than I give to any other truth-entailing property.

Consider an analogy. A die is about to be cast. Each of the natural

[13] The discussion of forgetting in section 10.2 provides one answer to the arguments to the contrary in Van Fraassen 1984 and 1995.

numbers from one to six has an equal prior probability (1/6) of being thrown. Exactly five propositions are inscribed on a rock:

e_1 A one will not be thrown.

e_2 A two will not be thrown.

e_3 A three will not be thrown.

e_4 A four will not be thrown.

e_5 A five will not be thrown.

The propositions inscribed on the rock are known to have been chosen at random; that a given proposition is inscribed there does not make it any more likely to be true. Say that a proposition is an *inscribed truth* if and only if it is true and inscribed on the rock; *pseudo-posterior* probabilities are the results of conditionalizing on the conjunction of all inscribed truths. Let p be the proposition that a six will be thrown. The prior probability of p is 1/6. If a one is thrown, then the inscribed truths are e_2, e_3, e_4, and e_5, so the pseudo-posterior probability of p is 1/2. By similar reasoning, the pseudo-posterior probability of p is 1/2 if any number between one and five is thrown. If a six is thrown, all the inscribed propositions are inscribed truths, and the pseudo-posterior probability of p is 1. Thus the pseudo-posterior probability of p is bound to be much higher than its prior probability. Its expected pseudo-posterior probability is $(5/6)1/2 + (1/6)1 = 7/12$. Pseudo-posterior probabilities are better informed than prior probabilities, because by definition they result from conditionalizing the latter on true and relevant information. All this is known in advance. Should we therefore revise our prior probabilities to bring them into line with our expected pseudo-posterior probabilities? We have no reason whatsoever to regard a six as any more likely to be thrown than any other number. The inscribed propositions embody a bias towards six. The bias could just as easily have been towards another number, quite independently of the result of the throw.

Moral: it is generally a mistake to try to align one's probabilities with what one knows about the results of conditionalizing them on truths with some given property. One instance of this mistake is to try to align our probabilities with what we know about the results of conditionalizing them on truths which we will know in the future. Although we may be made to suffer for the misalignment, it would not be rational to try to avert the suffering by changing our present beliefs. From our present perspective, the non-partitional structure of our future knowledge is a source of bias, similar in effect to forgetting although much subtler in its operation. Of course, we shall probably know more tomorrow, and it would be foolish then to disregard the new knowledge. But we cannot take advantage of the new knowledge in advance. We must cross that bridge when we come to it, and accept the consequences of our unfortunate epistemic situation with what composure we can find. Life is hard.

11
Assertion

11.1 RULES OF ASSERTION

We express and communicate our knowledge by making assertions. That by itself does not constitute a special relationship between knowing and asserting, for by making assertions we still express and communicate our beliefs when they fall short of knowledge. Indeed, assertion is the exterior analogue of judgement, which stands to belief as act to state. Nevertheless, there is a special relationship between knowing and asserting, if the argument of this chapter is correct. By analogy, there is also a special relationship between knowing and judging or believing. The relationship is a normative one.

Assertions are praised as true, informative, relevant, sincere, warranted, well phrased, or polite. They are criticized as false, uninformative, irrelevant, insincere, unwarranted, ill phrased, or rude. Sometimes they deserve such praise or criticism. If any respect in which performances of an act can deserve praise or criticism is a *norm* for that act, then the speech act of assertion has many norms. So has almost any act. Jumps can deserve praise as long or brave; they can deserve criticism as short or cowardly. But it is natural to suppose that some norms are more intimately connected to the nature of asserting than any norm is to the nature of jumping. One might suppose, for example, that someone who knowingly asserts a falsehood has thereby broken a rule of assertion, much as if he had broken a rule of a game; he has cheated. On this view, the speech act, like a game and unlike the act of jumping, is constituted by rules. Thus not all norms for assertion are on a par. Norms such as relevance, good phrasing, and politeness are just applications of more general cognitive or social norms to the specific act of assertion. Perhaps the norm of informativeness results from a more complex interaction between a general norm of cooperativeness and the nature of assertion as a source of information. But, on this view, not all norms for assertion derive from more general norms, otherwise nothing would differentiate it from other speech acts.

This chapter aims to identify the constitutive rule(s) of assertion, conceived by analogy with the rules of a game. That assertion has such rules is by no means obvious; perhaps assertion is more like a natural phenomenon than it seems. One way to find out is by supposing that it has such rules, in order to see where the hypothesis leads and what it explains. That will be done here. The hypothesis is not perfectly clear, of course, but we have at least a crude conception of constitutive rules, which we may refine as we elaborate the hypothesis. Although no attempt will be made here to *define* 'rule', some remarks on constitutive rules will focus the discussion.

Constitutive rules are not conventions. If it is a convention that one must ϕ, then it is contingent that one must ϕ; conventions are arbitrary and can be replaced by alternative conventions. In contrast, if it is a constitutive rule that one must ϕ, then it is necessary that one must ϕ. More precisely, a rule will count as constitutive of an act only if it is essential to that act: necessarily, the rule governs every performance of the act. This idealizes the case of games, for, in the ordinary sense of 'game', games such as tennis gradually change their rules over time without losing their identity; the constitutive role of the rules is qualified by that of causal continuity. Similarly, in the ordinary sense of 'language', natural languages such as English gradually change their rules over time without losing their identity. Nevertheless, in a technical sense of 'language' which the philosophy of language has found fruitful, the semantic, syntactic, and phonetic rules of a language are essential to it (Lewis 1975). The richer ordinary sense of 'language' introduces needless complications. Linguistic conventions and the consequent possibility of linguistic change can then be accommodated at a different point in the theory: a population which at one time has the convention of speaking a language L may later change to a convention of speaking a distinct language L*, constituted by slightly different rules. Likewise, in the present technical sense of 'speech act', the rules of a speech act are essential to it. A population which at one time has the convention of using a certain device to perform a speech act A may later change to a convention of using that device to perform a distinct speech act A*, governed by slightly different rules. 'Game' can receive a similar sense. Henceforth, 'rule' will mean constitutive rule.

Given a game G, one can ask 'What are the rules of G?'. Given an answer, one can ask the more ambitious question 'What are non-circular necessary and sufficient conditions for a population to play a game with those rules?'. Competent unphilosophical umpires know the answer to the former question but not to the latter. Given a language L, one can ask 'What are the rules of L?'. Given an answer, one can ask 'What are non-circular necessary and sufficient conditions for a population to speak a language with those rules?'. Given a speech act A, one can ask 'What are the rules of A?'. Given an answer, one can ask 'What are non-circular necessary and sufficient conditions for a population to perform a speech act with those rules?'. This chapter asks the former question about assertion, not the latter. It cannot wholly ignore the latter, for assertion is presented to us in the first instance as a speech act that we perform, whose rules are not obvious; in order to test the hypothesis that a given rule is a rule of assertion, we need some idea of the conditions for a population to perform a speech act with that rule, otherwise we could not tell whether we satisfy those conditions. Fortunately, we need much less than a full answer to the second question for these purposes. Our task is like that of articulating for the first time the rules of a traditional game that we play; that does not require a full philosophy of games.

Constitutive rules do not lay down necessary conditions for performing the constituted act. When one breaks a rule of a game, one does not thereby cease to be playing that game. When one breaks a rule of a language, one does not thereby cease to be speaking that language; speaking English ungrammatically is speaking English. Likewise, presumably, for a speech act: when one breaks a rule of assertion, one does not thereby fail to make an assertion. One is subject to criticism precisely because one has performed an act for which the rule is constitutive. Breaches of the rules of a game, language, or speech act may even be common. Nevertheless, some sensitivity to the difference—in both oneself and others—between conforming to the rule and breaking it presumably is a necessary condition of playing the game, speaking the language, or performing the speech act. The important task of elucidating the nature of this sensitivity will not be undertaken here.

The normativity of a constitutive rule is not moral or teleological. The criticism that one has broken a rule of a speech act is no more a moral criticism than is the criticism that one has broken a rule of a game or language. Although someone who knowingly asserts a falsehood may incur moral criticism, perhaps for having betrayed the hearers or inflicted false beliefs on them, such faults are made possible only by the specific nature of assertion, which is not itself constituted by moral norms. Cheating at a game is likewise not a morally neutral act, but it is made possible only by the non-moral rules which constitute the game. Nor is the criticism that one has broken a constitutive rule of an institution the criticism that one has used it in a way incompatible with its aim, whether the aim is internal or external. Consider a game, which might have the internal aim of scoring more goals than the opposition and the external aim of exercising players or entertaining spectators. Breaking the rules can serve both internal and external aims.

Conversely, lazy play can give away goals to the opposition, bore spectators, and fail to exercise players, without breaking the rules. Within the practice constituted by the rules, their authority does not require the backing of moral or teleological considerations.

What are the rules of assertion? An attractively simple suggestion is this. There is just one rule. Where C is a property of propositions, the rule says:

(The C rule) One must: assert *p* only if *p* has C.

In the imperative, assert *p* only if *p* has C. As used here, 'must' expresses the kind of obligation characteristic of constitutive rules. The rule is to be parsed as 'One must ((assert *p*) only if *p* has C)', with 'only if *p* has C' inside the scope of 'One must' but outside that of 'assert'. The rule unconditionally forbids this combination: one asserts *p* when *p* lacks C. The combination is possible, otherwise it would be pointless to forbid it. The condition that *p* has C may concern the content of the potential assertion (*p*), contextual features (for example, speaker and time), or both. The C rule is constitutive of the speech act: necessarily, assertion is a speech act A whose unique rule is 'One must: perform A with the content *p* only if *p* has C'. Furthermore, the envisaged account takes the C rule to be individuating: necessarily, assertion is the *unique* speech act A whose unique rule is the C rule. In mastering the speech act of assertion, one implicitly grasps the C rule, in whatever sense one implicitly grasps the rules of a game or language in mastering it. As already noted, this requires some sensitivity to the difference in both oneself and others between conforming to the rule and breaking it. All other norms for assertion are the joint outcome of the C rule and considerations not specific to assertion. If an assertion satisfies the rule, whatever derivative norms it violates, it is correct in a salient sense.[1] Call this account the C account, and any account of this form *simple.*

More complex accounts of assertion are conceivable. Some rules make some assertions obligatory; silence satisfies the C rule. There might be several rules of assertion. There might be none. Assertion might be wholly or partly constituted by a norm or norms whose normativity is not rule-like. Such a norm might be essentially comparative: mastery of the speech act would involve grasping a scale on which assertions could be assessed as better or worse than each other, but not

[1] In contrast, Dummett's notion of correctness for assertions is undifferentiated: 'to say one thing when there is no point in saying it rather than something more usual in those circumstances can be misleading if it prompts the hearers to suppose a point that does not exist; it, too, is therefore incorrect in the general sense' (1991: 168).

grasping a threshold for an assertion to be 'good enough'—that could be left to the discretion of individual speakers with particular purposes. Alternatively, assertion might be constituted only by non-normative features. Nevertheless, a simple account of assertion would be theoretically satisfying, if it worked. This chapter defends a simple account, shirking the examination of more complex accounts.

One obvious candidate to play the role of the property C is the truth of the content:

(The truth rule) One must: assert *p* only if *p* is true.

The truth rule forbids false assertions. It would be often broken, but so are many rules. The truth account—a simple account of assertion based on the truth rule—explains other norms as the joint outcome of the truth rule and considerations not specific to assertion. In particular, it explains epistemic norms as norms of evidence for truth: satisfying these secondary norms consists in having evidence that one satisfies the primary norm. Grice describes the category of Quality in his account of the rules of conversation in this vein: the supermaxim 'Try to make your contribution one that is true' leads to two more specific maxims, 'Do not say what you believe to be false' and 'Do not say that for which you lack adequate evidence' (Grice 1989: 27).

Unlike truth, other candidates to play the role of the property C are sensitive to the epistemic circumstances of the asserter. A speaker who satisfies such a condition will be described as having warrant to assert *p*, in a schematic sense of 'warrant'. On such views, the rule becomes:

(The warrant rule) One must: assert *p* only if one has warrant to assert *p*.

The warrant rule forbids unwarranted assertions. For any reasonable notion of warrant, a true assertion based only on a lucky guess will satisfy the truth rule without satisfying the warrant rule. Even so, versions of the warrant rule can be embedded in radically different simple accounts of assertion. On one kind of account, the content of an assertion consists in the condition for having warrant to make it, or perhaps it consists in that condition and the conditions for having warrant to make other structurally related assertions, for example of the negation of the assertion. This account reduces truth to some abstraction from warrant, and derives the norm of truth from the warrant rule. Such an account can be called *anti-realist*, although the term could equally well be applied to accounts of content in which truth plays no role at all.

The warrant rule can also be embedded in a different kind of account, on which having warrant to assert *p* amounts to knowing *p*. Then the warrant rule takes this form:

(The knowledge rule) One must: assert *p* only if one knows *p*.

The knowledge rule would be broken even more often than the truth rule, but so are many rules. The knowledge account—a simple account of assertion based on the knowledge rule—explains the norm of truth as a mere corollary of the knowledge rule: satisfying the former is a necessary but not sufficient condition of satisfying the latter. If one knows *p* then *p* is true. Nevertheless, this account in no way limits the transcendence of truth over warrant; still less does it make the former an abstraction from the latter. *Knows p* is not conceptually prior to *p*.[2] This account can be called *realist*. Given plausible connections between knowledge, belief, evidence, and truth, the knowledge account explains what is right about Grice's two more specific maxims of quality, 'Do not say what you believe to be false' and 'Do not say that for which you lack adequate evidence', as well as the supermaxim 'Try to make your contribution one that is true'.[3]

This chapter defends the knowledge account. The account can be roughly summarized in the slogan 'Only knowledge warrants assertion'. 'Warrant' is used here as a term of art, for that evidential property (if any) which plays the role of property C in the correct simple account of assertion. This use need not correspond exactly to that of 'warrant' in everyday English. It is not denied that false assertions are sometimes warranted in the everyday sense that they are sometimes reasonable; the claim is rather that the reasonableness of such assertions is explicable as the joint outcome of the knowledge rule and cognitive considerations not specific to assertion. Still, if the account is correct, ordinary speakers are implicitly sensitive to the knowledge rule, for they must have implicitly grasped it in mastering assertion. It is just that they need not use the word 'warrant' for that norm. Much of the evidence for the knowledge account comes from the ordinary practice of assertion.

[2] The knowledge account does not by itself guarantee realism. The term 'warranted assertibility' goes back to Dewey, who combines a kind of anti-realism with the identification of warranted assertibility and knowledge (1938: 7, 143) by means of a pragmatist conception of knowledge.

[3] This point is central to the case for the revised maxim of Quality 'Say only that which you know', in Gazdar 1979: 46–8 (if 'say' here can be read as 'assert'). Note that Grice's maxim of sincerity requires someone who asserts *p* not to believe ~*p*, not (as might be expected) to believe *p*. But if one asserts *p* while agnostic about *p*, one is insincere in a way that seems to flout conversational rules. Conversely, knowing *p* while unable to rid myself of what I recognize as an irrational belief in ~*p*, I might appropriately assert *p*, contrary to Grice's maxim but not the more obvious sincerity condition.

11.2 THE TRUTH ACCOUNT

It is somehow good to assert the true and bad to assert the false. Is that idea articulated by the truth account, the simple account of assertion based on the truth rule? This section argues that such an account is incorrect, and that its defects recommend the knowledge account.

One doubt about the truth account is that assertion is not the only speech act to aim at truth. For many speech acts A, normatively different from assertion and from each other, it is somehow good to perform A with a true content and bad to perform A with a false content. By definition, the truth account entails that the truth rule is individuating, in other words that assertion is the *unique* speech act A whose unique rule is 'Perform A with the content *p* only if *p* is true'. In this sense, the truth account claims that assertion is more intimately associated with the aim of truth than with any other speech act. But no basis is discernible for assigning this privilege to assertion in preference to all those other speech acts.

There is, for example, a speech act of *conjecturing p*, for which the evidential norms are more relaxed than they are for assertion. Although it is somehow good to conjecture the true and bad to conjecture the false, it is quite acceptable to conjecture *p*, but not to assert *p*, when *p* is merely more probable than not on one's evidence. In English, one can perform this speech act by using the words 'I conjecture' parenthetically, as in 'P, I conjecture' (compare Slote 1979: 182–7 on parenthetical uses of 'I believe'). Equally, there is a speech act of *swearing to p*, for which the evidential norms are more stringent than they are for assertion. Not only is it somehow good to swear to the true and bad to swear to the false, it is acceptable to swear to *p* only if one has grounds for unusual certainty about *p*, more than is required to assert *p*. In English, one can perform this speech act by using the words 'I swear' parenthetically, as in 'P, I swear'. What matters here is not the ordinary use of 'conjecture' and 'swear' but the possibility of speech acts of the kind described. Indeed, there is a whole range of possible speech acts, differing in their evidential norms but all in some sense aiming at the truth. Attempts to differentiate the speech acts and uphold the truth rule by adding rules about the gravity of breaches of it depart from the structure of a simple account; they also fail to meet an objection below to the truth account. Why should assertion be the only one of them to be a speech act A whose unique rule is 'Perform A with the content *p* only if *p* is true', as the truth account requires?

It might be held that, although asserting something is not always swearing to it, swearing to something is always asserting it. Swearing to *p* would be a solemn way of asserting *p*. This would not upset the argument. Conjecturing *p* is no way of asserting *p*. The evidential standard required for asserting would still be intermediate between those required for conjecturing and swearing-to. The question would remain: is that intermediate standard more intimately connected with the aim of truth than all the other standards are?

Simple accounts of assertion based on evidential rules face no such difficulty. They correspond to simple accounts of conjecturing and swearing-to based on evidential rules which require more and less respectively than does the rule of assertion. The speech acts are thereby differentiated from each other.

Although the preceding doubt about the truth account suggests (without showing) that the rule of assertion is evidential, it fails to indicate an appropriate standard of evidence. A stronger objection to the truth account will now be developed. It does cast light on the appropriate standard of evidence.

Assertion obviously has some kind of evidential norm. It is somehow better to make an assertion on the basis of adequate evidence than to make it without such a basis. Now assume the truth account, for an eventual *reductio ad absurdum*. Then the evidential norm is derivative from the truth rule. One ought to have evidence for one's assertions *because* they ought to be true.

The proposed derivation is simple. Its core is an inference from the premise that one must assert something only if it is true to the conclusion that one should assert it only if one has evidence that it is true. Since evidence that an assertion is true just is evidence for that assertion, the truth account implies that one should not make an assertion for which one lacks evidence. The underlying principle is quite general; it is not limited to assertion. The principle may be stated as a schema, with parentheses to indicate scope:

(1) If one must (ϕ only if *p* is true), then one should (ϕ only if one has evidence that *p* is true).

The transition from 'must' to 'should' represents the transition from what a rule forbids to what it provides a reason not to do. For example, if one must not bury people when they are not dead, then one should not bury them when one lacks evidence that they are dead. It is at best negligent to bury someone without evidence that he is dead, even if he is in fact dead. The proposed explanation of the evidential norm substitutes 'assert *p*' for 'ϕ' in (1). Clearly, there is much room for variation in the letter of (1) without violation of its spirit.

On a charitable reading of (1), the required weight of evidence for *p* will vary with the badness of ϕ-ing when p is false. One should take more care to avoid killing people than to avoid offending them, if the risks are equal in probability. The question is whether (1), so read, can explain the weight of evidence which we require speakers to have for their assertions in terms of the degree of badness which we attribute to making an untrue assertion. Is the former proportionate to the latter?

Consideration of lotteries suggests a negative answer. Suppose that you have bought a ticket in a very large lottery. Only one ticket wins. Although the draw has been held, the result has not yet been announced. In fact, your ticket did not win, but I have no inside information to that effect. On the merely probabilistic grounds that your ticket was only one of very many, I assert to you flat-out 'Your ticket did not win', without telling you my grounds. Intuitively, my grounds are quite inadequate for that outright unqualified assertion, even though one can construct the example to make its probability on my evidence as high as one likes, short of 1, by increasing the number of tickets in the lottery. You will still be entitled to feel some resentment when you later discover the merely probabilistic grounds for my assertion. I was representing myself to you as having a kind of authority to make the flat-out assertion which in reality I lacked. I was cheating.[4]

There is a special jocular tone in which it is quite acceptable to say '[Come off it—] Your ticket didn't win', but the tone signals that the speaker intends not to make a flat-out assertion. In the imagined example, I do not use that tone.

Can the fault in my assertion be explained by appeal to some version of (1)? The explanation would have to be that it is so bad to make an untrue assertion that one should not run even a minute risk of doing so. Is that plausible? We may well regard both honesty and the pursuit of truth as very serious matters, but it does not follow that we must regard every untrue assertion as a serious crime; the pursuit of truth would not get very far if we did. When we discover that we have inadvertently asserted something false on some casual matter, most of us are racked by no more guilt than we feel when we inadvertently tread on someone's toes. In the present case, let it be common knowledge between us that the result of the lottery will be announced within a few minutes, and that you care little whether your ticket wins. Thus the bad consequences of the falsity of my assertion which I risk inflicting on you—but do not

[4] Dudman 1992 uses the point to argue against the thesis that the assertibility of a conditional varies with the conditional probability of its consequent on its antecedent; Lowe 1995: 44 follows Dudman; contrast Edgington 1995: 287. For example, I am not entitled to assert to you 'If my ticket did not win then your ticket did not win', even though the probability of the consequent conditional on the antecedent is very high.

actually inflict, since my assertion is in fact true—amount to briefly having a false belief (if you believe my assertion) on a matter about which you care little. Ordinarily, we should not regard the fact that an action of mine involved a one-in-a-million risk of inflicting consequences of such limited badness on you as much cause for criticism. Yet you are entitled to insist that I was quite wrong to assert 'Your ticket did not win', for I had no authority to do so. That criticism of me does not derive from the kind of consideration embodied in (1). No assessment of the probability or gravity of untruth is even relevant to the criticism. The point is simply that, in making the assertion, I exceeded my evidential authority. In other cases, where untruth is less improbable or worse in its consequences if it does occur, the speaker is no doubt subject to *further* criticisms on those grounds, but they should not be allowed to obscure the possibility of criticizing speakers simply for exceeding their evidential authority.

Could a defender of the truth account explain what is wrong with my assertion by appeal to Gricean rules of conversation? The idea would be that my assertion was misleading because you, to whom I was speaking, were entitled to assume that the grounds on which I made it were not obviously already available to you, so you were entitled to assume that I had inside information about the result of the lottery. For making the assertion on grounds obviously already available to you might be held to violate one of the maxims of Quantity, 'Do not make your contribution more informative than is required' (Grice 1989: 26). However, if that Gricean point were the objection to my assertion, then the objection would extend to case (a), in which I assert 'Your ticket is almost certain not to have won', and the objection would not extend to case (b), in which I assert 'Your ticket did not win' but (unlike the previous cases) it is not obvious that you know how many tickets other than your own have been sold. For in case (a), parallel Gricean reasoning would indicate that you are entitled to assume that the grounds on which I made my assertion were not obviously already available to you, and therefore that you are entitled to assume that I had inside information about the result of the lottery—for example, evidence that it was almost certain to have been rigged in favour of someone else. In case (b), my grounds for the assertion—the number of tickets sold—are not obviously already available to you, so the assumption to that effect, which the argument supposes you to be entitled to make, is true, and the Gricean objection lapses. In fact, however, the problem behaves in the opposite way to that predicted by the Gricean explanation. It does not extend to case (a), in which the worst to be said of my assertion is that it is banal and unkind. The problem does extend to case (b), in which you are still entitled to feel resentment at the merely probabilistic grounds for my assertion. Probabilistic evidence warrants only an assertion that something is probable.

A further problem for the Gricean explanation is that I should be able to remove the objection to my assertion by explicitly cancelling the supposed conversational implicature. I am not. I have no more evidential authority to assert 'Your ticket did not win, but I do not mean to imply that I have inside information' than I have to assert the plain 'Your ticket did not win'. The criticisms here are not of Grice's theory of conversational implicature itself but only of an over-enthusiastic application of it.

A different defence of the truth account appeals to the point that, for each ticket, I have a similar basis for asserting that it did not win. If I make all those assertions, I shall have asserted something false. But how does that explain what is wrong with making any one of them, granted the truth rule? Consider an analogy. I am faced with an enormous pile of chocolates. I know that exactly one of them is contaminated and will make me sick; alas, I cannot tell them apart. I have a strong desire to eat a chocolate. I can quite reasonably eat just one, since it is almost certain not to be contaminated, even though, for each chocolate, I have a similar reason for eating it, and if I eat all the chocolates, I shall eat the contaminated one, and my sickness will be overdetermined. No plausible principle of universalizability implies that, in the circumstances, any reason for taking one chocolate is a reason for taking them all; the most to be implied is that, in the circumstances, any reason for taking one chocolate is a reason for taking any other chocolate instead. The truth account does not supply the resources to rule out the possibility that there is adequate evidence for each of the assertions '*t* did not win' but not for their conjunction. If each conjunct is true then the conjunction is also true, of course, but it does not automatically follow that the same goes for adequate evidence of truth. Although the principle that entitlement to assert each conjunct implies entitlement to assert the conjunction may be independently plausible, the truth account cannot explain it.

It is not even essential to the lottery case that each ticket should have an equal chance of winning. Consider a variant lottery in which each ticket is assigned a publicly known weight proportional to its probability of winning, and your ticket has a somewhat lower weight than the others. I am still not entitled to assert that your ticket will not win, even though my evidence that it will not win is now better than for any other ticket. Alternatively, the lottery might even be one in which there was probably *no* winning ticket (DeRose 1996).

It might finally be protested that if I lacked warrant to assert 'Your ticket did not win', then we lack warrant to make most of our ordinary assertions, because few of them are quite certain. Of course, it follows that I had warrant only given the anti-sceptical premise that we do have warrant to make most of our ordinary assertions. The protest simply assumes that no other account of assertion can discriminate between 'Your ticket did not win' and most of our ordinary assertions. That assumption needs testing; it is rejected in the next section. In any case, for whatever reasons, probabilistic bases are ordinarily taken to be inadequate for assertion.

The truth account does not explain something that it is committed to explaining: the evidential norms for assertion. It should therefore be rejected. A speech act genuinely based on the truth rule would be more like the act of saying; one can say something without asserting it, for example, in guessing the answers to a quiz (Unger 1975: 267 credits the point to Harman). Assertion itself seems to be governed by a non-derivative evidential rule, which my assertion in the lottery case broke; I was cheating.

One possible explanation is this. The rule of assertion is the knowledge rule; one must not assert *p* unless one knows *p*. In the lottery case, it is intuitively clear, given the nature of my evidence, that I did not know that your ticket did not win.[5] Thus my assertion violated the rule of assertion. After all, the natural way for you to articulate the criticism that I lacked evidential authority for my assertion is by saying 'But you didn't *know* that my ticket hadn't won!'. This argument will be developed in the next section.

11.3 THE KNOWLEDGE ACCOUNT

One may lack the evidential authority to assert a proposition about a lottery, even though the proposition is very highly probable on one's evidence. It will now be argued that the underlying phenomenon is general to assertions about any subject matter.

Let *p* be a proposition whose truth value is known to an expert but about which you have no evidence. The expert holds a lottery. There are

[5] Arguing for an externalist view of knowledge, Armstrong denies that we have knowledge in the lottery case (1973: 185–8); in defending an internalist view of knowledge from Armstrong's argument, BonJour concedes the absence of knowledge but proposes an internalist explanation of it (1985: 56). See also Harman 1968, Dretske 1981a: 99–102, Craig 1990a: 98–103, DeRose 1996, and Lewis 1996.

a million tickets, of which you have one. However, she does not announce the number of the winning ticket; she merely hands each participant a slip of paper. If your ticket won, the true member of the pair {*p*, ~*p*} is written on your slip; if your ticket lost, the false member of the pair is written there. There is no doubt that this is the arrangement. You are not in a position to confer with other participants. Suppose that ~*p* turns out to be written on your slip. On your evidence, there is a probability of one in a million that your ticket won and ~*p* is true, and a probability of 999,999 in a million that your ticket lost and ~*p* is false. Thus, if you assert *p*, the probability on your evidence that your assertion is true is 999,999 in a million. Intuitively, however, you are not entitled to assert *p* outright. That intuition can be supported. On your evidence, you can certainly assert the biconditional linking *p* and 'My ticket did not win'; by hypothesis, it is not in doubt. If you assert *p*, you are therefore surely in a position to detach, and assert 'My ticket did not win'. But you are not entitled to assert that, for your only evidence is that your ticket was one in a million. That ~*p* rather than *p* was written on your slip tells you nothing, for you have no independent evidence for or against those propositions. Thus you are not entitled to assert *p*, even though it has a probability on your evidence of 999,999 in a million.

In the preceding example, *p* could be *any* assertion about which you happen to have no evidence. Indeed, even if you have probabilistic evidence that tends to support ~*p*, the number of tickets in the lottery can be made so large that your probabilistic evidence from the lottery for *p* will overwhelm your other evidence against *p*. Thus the argument indicates that, for almost any kind of proposition at all, very high probability on one's evidence does not imply assertibility. The propositions not covered by the argument are those for which one is bound to have independent non-probabilistic evidence, for example, 'I exist': but they are not plausible candidates for assertion on a merely probabilistic basis. The obvious moral is that one is *never* warranted in asserting a proposition by its probability (short of 1) alone. What matters in the original lottery case is not the subject matter of the assertion but the probabilistic basis on which it was made.

To say that no probability short of 1 warrants assertion is not yet to say that only knowledge warrants assertion. Some non-deductive forms of inference might be held sometimes to warrant assertion non-probabilistically without providing knowledge; an example is inference to the best explanation. It is hard to see how inference to the best explanation could ever generate numerical probabilities, but even if it does lead to conclusions of high probability short of 1, it would not warrant assertion *in virtue* of doing so. The implication is that one might have warrant to assert the conclusion of an inference to the best explanation, even though one lacked warrant to assert an equally probable proposition whose high probability had a different basis; no inference to the best explanation provided the probabilistic evidence that your ticket did not win. Such a view is consistent, but is it plausible? If one has warrant to assert a proposition of probability less than 1 on one's evidence, then in some lottery case one lacks warrant to assert a proposition—perhaps the very same proposition—of higher probability on one's evidence.

Assume, plausibly, that if *p* is less probable than *q* on one's evidence, and one has warrant to assert *p*, then one has warrant to assert *q*. Given our intuitions about lotteries, it follows that one never has warrant to make assertions of probability less than 1 on one's evidence. This conclusion might appear to be a sceptical one, even a reductio ad absurdum. For it is easy to suppose that almost all our ordinary empirical assertions are of probability less than 1 (for example, Edgington 1995: 287). But what kind of probability is in question? If it is objective probability, then the problem affects only assertions about the future, for only they have an objective probability other than 1 or 0. But objective probability is too objective to warrant assertion: of two past tense assertions whose objective probability is 1, I may have excellent evidence for one and none for the other. Equally, subjective probability (degree of belief) is too subjective to warrant assertion: I do not gain warrant to assert that I am Napoleon merely from my baseless conviction that I am Napoleon, even if my conviction is so dogmatic that the assertion has subjective probability 1. If any probability warrants assertion, it is probability *on one's evidence*.

What is one's evidence? The simple answer defended in Chapters 9 and 10 is available to the knowledge account: one's evidence is just what one knows. We could additionally argue for it from the knowledge account of assertion, given the not wholly uncontroversial premise that one's evidence consists of just those propositions which the rules of assertion permit one to assert outright. The equation of knowledge with evidence was not assumed in the earlier discussion of evidence for assertions. For present purposes, it would not matter if it were considered to sharpen the prior notion rather than merely elucidating it, for either way the result is *a* tenable notion of evidence. Without making any substantive assumptions about the conditions for knowledge, this view makes it trivial that if one knows *p*, then the probability of *p* on one's evidence is 1. This does not imply that no discovery could shake one's confidence in *p*, for discoveries can undermine knowledge. Nor does it imply that one would in practice bet one's life against a penny on *p*; that test defines no useful notion of probability (let *p* be a moderately complicated tautology). The standard of probability 1 on one's evidence is no more demanding than the standard of knowledge.

The denial of knowledge in the lottery case might also be feared to have sceptical implications, on the grounds that virtually all our empirical knowledge has a probabilistic basis. For example, our perceptual processes are subject to random error. However, one must distinguish between causal and evidential senses of the word 'basis'. The causal connection between the environment and our perceptual beliefs about it is no doubt probabilistic, but it does not follow that those beliefs rest on probabilistic evidence. On the view above of evidence, when they constitute knowledge, they are part of our evidence. Moreover, they may constitute knowledge simply because perceiving counts as a way of knowing; that would fit the role of knowledge as evidence (see section 1.4). I certainly did not *perceive* that your ticket did not win. There is no valid argument from the denial of knowledge in the lottery case to its denial in perceptual and other cases in which we ordinarily take ourselves to know. The knowledge rule provides a better explanation of the inadequacy of probabilistic grounds for assertion than do accounts on which something less than knowledge warrants assertion.

Conversational patterns confirm the knowledge account.[6] Consider a standard response to an assertion, the question 'How do you know?'. The question presupposes that it has an answer, that somehow you do know. If not only knowledge warrants assertion, what makes that presupposition legitimate? The question 'Where did you read that?' is not normally appropriate in response to an assertion, because someone who asserts *p* is not usually committed to having read *p* somewhere. But 'How do you know?' is normally appropriate. Of course, it is silly to ask 'How do you know?' when the questioner obviously knows as well as the asserter how the latter knows, for example, when someone has

[6] Much of the relevant evidence is marshalled by Unger 1975: 250–65 and Slote 1979. The thesis defended by Unger and Slote is that, in asserting *p*, one *represents* oneself as knowing *p*; see also DeRose 1991: 598–605. These authors say little about the general notion of representation, which this chapter scarcely employs. The knowledge account subsumes the Unger–Slote thesis under more general principles. In doing anything for which authority is required (for example, issuing orders), one represents oneself as having the authority to do it. To have the (epistemic) authority to assert *p* is to know *p*. The Unger-Slote thesis follows. Lloyd Humberstone suggests that Max Black may have originated this talk of representing oneself: 'In order to use the English language correctly, one has to learn that to pronounce the sentence "Oysters are edible" in a certain tone of voice is to *represent oneself* as knowing, or believing, or at least not disbelieving what is being said. (To write a cheque is to represent oneself as having money in the bank to honour the cheque)' (Black 1952: 31). G. E. Moore makes a related claim: 'by asserting *p* positively you *imply*, though you don't assert, that you know that *p*' (1962: 277; see also 1912: 125).

said 'I want to go home'. And the questioner does not always *believe* the presupposition of the question, for it is sometimes (not always) intended as a challenge to the assertion. Nevertheless, it is an *implicit* challenge: the questioner politely grants that the asserter does know *p*, and merely asks how, perhaps suspecting that there is no answer to the question. If not only knowledge warranted assertion, the absence of an answer would not imply the absence of a warrant; why should the question constitute even an implicit challenge? The hypothesis that only knowledge warrants assertion makes good sense of the phenomenon.

A less standard and more aggressive response to an assertion is the question 'Do you know that?'. Its aggressiveness is easy to understand on the hypothesis that only knowledge warrants assertion, for then what it calls into question is the asserter's warrant for the assertion. On the hypothesis that not only knowledge warrants assertion, the aggressiveness of the question is hard to understand, for the asserter might truthfully answer 'No' and still have warrant for the assertion.

A related argument starts from a version of Moore's paradox, with 'know' in place of 'believe' (Moore 1962: 277; Unger 1975: 256–60; Jones 1991). Something is wrong with any assertion of the form 'A and I do not know that A', even though such assertions would often be true if made. What is wrong can easily be understood on the hypothesis that only knowledge warrants assertion. For then to have warrant to assert the conjunction 'A and I do not know A' is to know that A and one does not know A. But one cannot know that A and one does not know A. One knows the conjunction only if one knows each conjunct, and therefore knows that A (the first conjunct); yet one knows the conjunction only if it is true, so only if each conjunct is true, so only if one does not know that A (the second conjunct); thus the assumption that one knows the conjunction that A and one does not know that A yields a contradiction. Given that only knowledge warrants assertion, one therefore cannot have warrant to assert 'A and I do not know that A'.[7] In contrast, the hypothesis that not only knowledge warrants assertion makes it hard to understand what is wrong with an assertion of that form. One often has good evidence that A whilst knowing for sure that one does not know that A; in such cases one has good evidence short of knowledge for the conjunction that A and one does not know that A. If good evidence short of knowledge warranted assertion, one would have warrant to assert 'A and I do not know that A': but one has not. For

[7] Hintikka's explanation of the defectiveness of such assertions by their epistemic indefensibility assumes that assertions ought to be epistemically defensible in his sense (1962: 78–102). That assumption is unmotivated unless only knowledge warrants assertion. See Chapter 12 for more on this kind of argument.

example, I have excellent evidence for the conjunction that your ticket did not win and I do not know that your ticket did not win. If such evidence warranted assertion, I should have warrant to assert 'Your ticket did not win and I do not know that your ticket did not win'.

Given that knowing entails believing, a similar explanation reveals what is wrong with any assertion of the more familiar Moorean form 'A and I do not believe that A', for one knows the first conjunct only if the second conjunct is false (Sorensen 1988: 15–56 has a more general account of Moorean paradoxes along similar lines).

Naturally, these arguments apply only to utterances of the conjunction within a single context. If the contextual standards for knowledge are raised between the utterance of the first conjunct and that of the second, the assertion might be acceptable. Its unacceptability within a single context must still be explained.

Knowledge is not even a cancellable implication of assertion (Slote 1979: 179). For if the implication could be cancelled, the second conjunct 'I do not know that A' would cancel it, and it would be acceptable to assert the conjunction; but it is not acceptable.

One might fear that such arguments would prove too much. After all, something is wrong even with the assertion 'A and I cannot be certain that A'. Does that not suggest that only something *more* than knowledge warrants assertion? What seems to be at work here is a reluctance to allow the contextually set standards for knowledge and certainty to diverge. Many people are not very happy to say things like 'She knew that A, but she could not be certain that A'. However, we can to some extent effect such a separation, and then assertibility goes with knowledge, not with the highest possible standards of certainty. For example, one may have warrant to assert 'A and by Descartes's standards I cannot be absolutely certain that A', where the reference to Descartes holds those standards apart from the present context. Again, it would often be inappropriate to respond to the assertion 'A' by asking 'How can you be so certain that A?'. The word 'so' flags the invocation of unusually high standards of certainty. By ordinary standards you may have had warrant to assert that A even if you could not be *so* certain that A.

The putative connections between knowledge, assertion and certainty contain an obvious sceptical threat (elaborated in Unger 1975). One response is to permit contextual variation in epistemic standards: in effect, 'know' would express different contents in different contexts, as a result of either variation in meaning or an invariant indexical meaning (DeRose 1995 and Lewis 1996 are recent examples). If so, 'assert' will express correspondingly different contents. The contents will nevertheless have enough in common to be appropriately discussed together, as is standard in contextualist epistemology. The present account permits such contextual variation, but, as argued in Chapters 7, 8, and 9, it can resist the concession that sceptical arguments create a context in which sceptical utterances express truths.[8]

Considerable evidence has emerged that our ordinary linguistic practice acknowledges the knowledge rule. Certain phenomena are nevertheless likely to be adduced as counter-evidence. The next section considers such objections.

11.4 OBJECTIONS TO THE KNOWLEDGE ACCOUNT, AND REPLIES

That false beliefs are often reasonable is a commonplace. The account of evidence in Chapters 9 and 10 allows a false proposition to be very highly probable on one's evidence, even though one's evidence itself is true. Evidence can be misleading. When one reasonably but falsely believes *p*, is it not reasonable to *assert p*, even though one does not know *p*? If so, what becomes of the claim that only knowledge warrants assertion?

On some views, it is sometimes reasonable to believe *p*, even though one knows that one does not know *p*. For example, it is reasonable for me to believe that I shall not be run over by a bus tomorrow, even though I know that I do not know that I shall not be run over by a bus tomorrow (Slote 1979: 180; I am not confined in a bed, lost in a jungle, or the like). Such cases do not threaten the hypothesis that only knowledge warrants assertion, for they are ones in which, intuitively, assertion is not warranted. It would be foolish of me baldly to assert that I shall not be knocked down by a bus tomorrow; it would invite the objection 'You don't know that'. As in the lottery case, I should assert no more than that it is very unlikely that I shall be knocked down by a bus tomorrow. Such cases support the hypothesis that only knowledge warrants assertion.

It is plausible, nevertheless, that occurrently believing *p* stands to asserting *p* as the inner stands to the outer. If so, the knowledge rule for assertion corresponds to the norm that one should believe *p* only if one

[8] See Hambourger 1987 for a form of contextualism applied to both 'know' and 'assert'. However, Hambourger's claimed identity of the evidential standards for asserting 'A' and 'I know that A' must be rejected here, for I may know that A without knowing that I know that A (see Chapter 5). His claim would make 'A' co-assertible with 'I know that I know that I know that . . . A'.

knows p (see also section 1.5). Given that norm, it is not reasonable to believe p when one knows that one does not know p. If one knows that what one knows is only that p is very probable, then what it is reasonable for one to believe is only that p is very probable. For example, I should not believe that my ticket will not win the lottery. Outright belief in this sense requires more than a high subjective probability as determined by betting ratios (see section 4.4). On this analogy between assertion and belief, the knowledge rule for assertion does not correspond to an identification of reasonable belief with knowledge (contrast Wright 1996: 935). The rule makes knowledge the condition for permissible assertion, not for reasonable assertion. One may reasonably do something impermissible because one reasonably but falsely believes it to be permissible. In particular, one may reasonably assert p, even though one does not know p, because it is very probable on one's evidence that one knows p. In the same circumstances, one may reasonably but impermissibly believe p without knowing p. That possibility is consistent with the equation of evidence with knowledge.[9]

Sometimes one knows that one does not know p, but the urgency of the situation requires one to assert p anyway. I shout 'That is your train', knowing that I do not know that it is, because it probably is and you have only moments to catch it. Such cases do not show that the knowledge rule is not the rule of assertion. They merely show that it can be overriden by other norms not specific to assertion. The other norms do not give me warrant to assert p, for to have such warrant is to satisfy the rule of assertion. Similarly, when I am speaking a foreign language, the urgency of the situation may require me to speak ungrammatically, because it would take me too long to work out the correct grammatical form for what I want to say; it does not follow that my utterance satisfied the rules of grammar in that context.

In other cases, one reasonably but falsely believes p, and is in no position to know that one does not know p (see also section 8.2). One cannot discriminate between one's actual circumstances and circumstances

[9] We can model the possibility in epistemic logic (section 10.4) using worlds $w_0, \ldots, w_n$, where all worlds are epistemically accessible from w_0, only w_i is accessible from w_i ($1 \leq i \leq n$), the prior probability distribution is uniform and p is true at all worlds except w_0. Then Kp is also true at all worlds except w_0. Since all worlds are accessible from w_0, one's evidence (what one knows) at w_0 is true at all worlds, so the probability of Kp on one's evidence at w_0 is its prior probability, which is $n/(n+1)$. Thus although p and Kp are false at w_0, both can have probabilities on one's evidence there arbitrarily close to 1, for suitably large n. A necessary condition for such examples is the failure of the S5 principle ~Kp → K~Kp, for that principle gives Kp probability 0 on one's evidence wherever it is false. The failure of the S5 principle is also sufficient for such examples on a suitable probability distribution, since it means that Kp is true at some worlds accessible from a world at which Kp is false; the distribution can be weighted in favour of those worlds.

in which one would know *p*. For example, it is winter, and it looks exactly as it would if there were snow outside, but in fact that white stuff is not snow but foam put there by a film crew of whose existence I have no idea. I do not know that there is snow outside, because there is no snow outside, but it is quite reasonable for me to believe not just that there is snow outside but that I know that there is; for me, it is to all appearances a banal case of perceptual knowledge. Surely it is then reasonable for me to assert that there is snow outside.

The case is quite consistent with the knowledge account. Indeed, if I am entitled to assume that knowledge warrants assertion, then, since it is reasonable for me to believe that I know that there is snow outside, it is reasonable for me to believe that I have warrant to assert that there is snow outside. If it is reasonable for me to believe that I have warrant to assert that there is snow outside, then, other things being equal, it is reasonable for me to assert that there is snow outside. Thus the knowledge account can explain the reasonableness of the assertion. However, granted that it is reasonable for me to believe that I have warrant to assert *p*, it does not follow that I do have warrant to assert *p*. The term 'warrant' has been reserved for the property C in the rule C of assertion. There may be other evidential norms for assertion, if they can be derived from the knowledge rule and considerations not specific to assertion. The reasonableness of asserting *p* when one reasonably believes that one knows *p* has just been derived in exactly that way.

One can think of the knowledge rule as giving the condition on which a speaker has the *authority* to make an assertion. Thus asserting *p* without knowing *p* is doing something without having the authority to do it, like giving someone a command without having the authority to do so. Characteristic standards of authority thus play a constitutive role in the speech act of assertion, as they do in other institutions. The distinction between having warrant to assert *p* and reasonably believing oneself to have such warrant becomes a special case of the distinction between having the authority to do something and reasonably believing oneself to have that authority. Someone who does not know *p* lacks the authority to assert *p*, and therefore cannot pass that authority on to me by asserting *p*, no matter how plausibly he gives me the impression that he has done so. Although there are special cases in which someone comes to know *p* by hearing someone who does not know *p* assert *p* (Lackey 1999), the normal procedure by which the hearer comes to know *p* requires the speaker to know *p* too.

The assimilation of warrant to authority is misleading in one respect. Authority, even intellectual authority, usually extends over an area; it is not confined to a single proposition. In the present sense, testimony can give one the authority to assert *p*, even if one is pitifully ignorant about neighbouring questions, and no extent of knowledge about neighbouring questions can give one the authority to assert *p* if one happens to be mistaken on that single point.

We are not always in a position to know whether we know *p* (Chapters 5 and 8). The knowledge account therefore implies that we are not always in a position to know whether we have warrant to assert *p*. We are liable to error and ignorance about warrant, just as we are about almost everything else (Chapter 4). This view of warranted assertibility is in sharp contrast with its treatment in anti-realist theories of meaning to which the notion of the assertibility conditions of sentences is crucial. Such theories characteristically assume that one has no difficulty in knowing whether one has warrant to assert *p*. Independently of the knowledge account, there is reason to doubt that there could be a norm of the kind postulated by anti-realist theories (section 4.8 and Chapter 8).[10]

The knowledge account may seem to imply that speakers should always be at great pains to verify a proposition before asserting it. The wide variety of situations in which speakers go to no such pains may therefore seem to threaten the knowledge account: consider a lively seminar discussion, or gossip. To rule that speakers are not making genuine assertions in such situations would be to trivialize the account. In natural languages, the default use of declarative sentences is to make assertions, and the situations at issue are not special enough to cancel the default. Rather, the point is that the knowledge account does not imply that asserting *p* without knowing *p* is a terrible crime. We are often quite relaxed about breaches of the rules of a game which we are playing. If the most flagrant and the most serious breaches are penalized, the rest may do little harm. In some sports, it is said that some rules are being breached most of the time. Similarly, many of the utterances in an ordinary conversation are syntactically ill formed even by

[10] There are related difficulties for the account of assertion in Brandom 1983, where a central normative role is played by a notion of justification for which it is claimed that 'a justification is whatever the community treats as one—whatever its members will let assertors get away with' (644) and knowledge is a matter of 'an appropriately justified true belief' (647). The community is neither omniscient nor infallible in judging when true belief amounts to knowledge. The usual problems would attend an appeal to epistemically ideal counterfactual conditions. Brandom's emphasis on the ability to articulate inferential justifications of one's assertions in response to challenges (641–2) also seems to overestimate the link between warranted assertion and a smooth tongue. See also Brandom 1994 for his account, on which 'assertions . . . have the significance of claims to knowledge' (1994: 202). However, Brandom's conception of knowledge as a hybrid deontic status (justified true belief or the like) clashes with the account in Chapter 1. It is also not clear that his approach makes the right predictions about lottery cases.

the standards of the speaker's own idiolect, for example as a result of intentional or unintentional changes of direction in mid-sentence. Breaches of the rules are more serious in writing than in speech; that applies to the rule of assertion, too. When assertions come cheap, it is not because the knowledge rule is no longer in force, but because violations of the rule have ceased to matter so much.

To be relaxed in applying a rule is not to replace it by a different rule. Even in a lively seminar discussion, or gossip, the knowledge rule does not give way to a rule of reasonable belief. For example, even in gossip, it would be cheating to assert 'Mr Jones won nothing in the lottery again last week' merely on the basis of its high probability. Similarly, if I overhear an expert logician in a room full of people say 'A flaw has just been found in the proof of the main theorem in his last paper' when it is 99 per cent probable that the person whom he is demonstrating is Professor X, I may form a reasonable belief that a flaw has just been found in the proof of the main theorem of Professor X's last paper, but, even in a lively seminar discussion, it would be cheating for me to answer someone who bases an objection to my views on that theorem by asserting, without qualification, 'A flaw has been found in the proof of that theorem'. Such assertions are unacceptable because the speaker knows that he lacks the requisite knowledge, even though he has a reasonable belief. When we are relaxed in applying the rule, we feel entitled to assert *p* whenever we are not confident that we do not know *p*. We still try to obey the knowledge rule, but we do not try very hard.

In debate, we are often willing to assert *p* when we do not expect to persuade our interlocutors of *p*. However, knowing *p* is quite consistent with being unable to persuade other people of *p*. Knowledge often depends on good judgement, the speaker may have better judgement than the hearer, and most speakers value their own judgement more highly than they know their hearers do.

Some people use the locution 'I assert that . . .' only when they cannot supply compelling grounds; the implied contrast is with 'I can prove that . . .' or the like. For the reason just given, they are not conceding that they do not know. The simplest analysis of what one does in uttering the syntactically declarative sentence 'I assert that A' is that one asserts that A by asserting that one asserts that A—just as, in uttering 'I promise to ϕ', one promises to ϕ by asserting that one promises to ϕ (Lemmon 1962; Hedenius 1963; Heal 1974; Lewis 1983: 224, from Lewis 1970; Ginet 1979; for a different but related view, Recanati 1987: 169–75). On that view, one obviously knows that one asserts that A, and therefore is warranted in asserting that one asserts that A. This may help to distract attention from the more problematic question: is one warranted in asserting that A? One dodges that question by focusing one's hearers' attention on the less contentious assertion.

11.5 THE BK AND RBK ACCOUNTS

I may believe on good evidence that your lottery ticket did not win; I am not warranted in asserting that it did not win. I may believe on good evidence that I shall not be knocked down by a bus tomorrow; I am not warranted in asserting that I shall not be knocked down by a bus tomorrow. Neither belief nor belief on good evidence warrants assertion. Nevertheless, it might still be thought that some false assertions are warranted in the technical sense that they obey the rule of assertion. One proposal along such lines is that the rule for assertion is this:

> (The BK rule) One must: assert *p* only if one believes that one knows *p*.

(Thijsse forthcoming, citing similar views from Lenzen 1980). What one believes oneself to know need not be true. Can the BK account explain the phenomena?

The BK account can explain many of the conversational phenomena that were used as evidence for the knowledge account by adapting the latter's explanations to its own use. For example, I can follow the proof which shows that I cannot know the conjunction that A and I do not know that A, and should therefore refrain from believing that I know that A and I do not know that A. If I do so refrain, then the assertion 'A and I do not know that A' would violate the BK rule. Similarly, if I am committed to believing that I know by my assertion, then the challenge 'How do you know?' has an obvious relevance.

One problem for the BK account is that my belief that I know *p* may be as irrational as any other belief. The BK account's analysis of the modified Moorean sentence depends on the assumption that if 'B' is inconsistent then 'I believe that B' is inconsistent, which is invalid for subjects who are logically capable of irrationality. Suppose that I have an irrational belief that I know that G. E. Moore was a serial killer. On the BK account, my assertion 'G. E. Moore was a serial killer' satisfies the rule of assertion. Neither its falsity nor its irrational basis constitutes a breach of the BK rule. So far, nothing is wrong with the assertion itself. Plenty is wrong with the asserter, for I have a completely irrational belief, but that is another matter. Although I have obeyed the BK rule only by expressing an irrational belief, the BK account lacks the resources to explain why that is a fault in the assertion itself. Defenders of the BK account cannot deny that we distinguish faults in the assertion from faults in the asserter. If I am asked 'Do you really have the belief that G. E. Moore was a serial killer?' (a question about me, not about G. E. Moore) then, in the circumstances, I ought to answer 'Yes', which is to assert that I have the belief that G. E. Moore was a serial killer; the assertion itself is quite in order, even though its being so depends on my irrational belief that G. E. Moore was a serial killer. The fault there is clearly in the asserter, not in the assertion. Since the BK account cannot explain why we regard the assertion 'G. E. Moore was a serial killer', not just its assertor, as faulty, it should be rejected. In contrast, the knowledge account has no difficulty in explaining what is wrong with the assertion, for it breaks the knowledge rule.

On an obvious revision of the BK account, the rule for assertion is this:

> (The RBK rule) One must: assert *p* only if one rationally believes that one knows *p*.

The added condition of rationality both improves the analysis of modified Moorean sentences and eliminates the counterintuitive consequences above. Nevertheless, all is not well with the RBK account. One problem concerns conjunctive assertions. Consider a complicated paradox, in which a contradiction is deduced from a very large number of premises p_1, . . ., p_n. For each number *i*, p_i seems intuitively obvious; indeed, it seems intuitively obvious that we know p_i. Even on reflection, we are quite unsure which premise to blame for the contradiction. Suppose also that it is unlikely that more than one of the premises is false; each premise seems to have a quite different basis from the others, so that its falsity would be unlikely to infect them. Then we might easily, for each number *i*, rationally believe ourselves to know p_i. For each *i*, on the RBK account, we therefore have warrant to assert p_i. Nevertheless, we know that the conjunction of p_1, . . ., p_n is false, because it entails a contradiction; thus it is not rational to believe ourselves to know the conjunction, so, on the RBK account, we lack warrant to assert it. Warrant to assert would not be closed under conjunction. This consequence of the RBK account is disturbing, but not clearly absurd.[11]

[11] Arguably, the knowledge rule does not preserve warrant to assert under conjunction either, since one might know some propositions without having entertained their conjunction. However, this would be an insignificant failure of closure compared with that in the text, which remains even when the subject has carefully considered the conjunction.

The RBK account shares a simpler problem with the BK account; the analogue for falsity of the latter's problem about irrationality. Suppose that I rationally believe myself to know that there is snow outside; in fact, there is no snow outside. On the BK and RBK accounts, my assertion 'There is snow outside' satisfies the rule of assertion. Yet something is wrong with my assertion; neither the BK nor the RBK account implies that it is. They can allow that something is wrong with my belief that I know that there is snow outside, for it is false, but that is another matter. The BK and RBK accounts lack the resources to explain why we regard the false assertion itself, not just the asserter, as faulty.

A further objection to the BK and RBK accounts, an obvious methodological one, is that they are less simple than the account based on the knowledge rule. Their adoption might be reasonable if the latter were refuted, but it has not been. The onus of proof is on the BK and RBK accounts. After all, when I assert *p*, why should it matter whether I rationally believe myself to know *p* if I am not required to know *p*? Of course, the truth account is even simpler than the knowledge account, so the onus of proof is on the latter against the former; but it has already been discharged.

A final objection to the RBK account is that it makes it too easy for someone who lacks the authority to assert *p* to confer that authority on someone else. For even if one knows ~*p*, one might create sufficiently misleading appearances to make others reasonably but falsely believe themselves to know *p*. Intuitively, such a trick confers only the appearance, not the reality, of the authority to assert *p*; according to the RBK account, it confers genuine authority. If a truth requirement were added to the RBK rule ('One must: assert *p* only if one rationally and truly believes that one knows *p*'), it would require knowledge after all. One should adopt the simpler knowledge account instead.

One possible motivation for belief-based accounts of assertion (hinted at by Thijsse) is the idea that what warrants assertion should be a mental state of the asserter. On a common view, believing and reasonably believing oneself to know *p* are mental states, while knowing *p* is not. However, it was argued in Chapter 1 that knowing *p* *is* a mental state, of an externalist kind. Indeed, the combination of that idea with the idea that falsity is a fault in the assertion itself, so that what warrants asserting *p* entails *p*, implies that what warrants asserting *p* is a mental state which entails *p*. Knowing *p* is the best candidate for such a state. On a more internalist conception of mental states, this question would become more pressing: why should what warrants assertion be a mental state of the asserter? One bad answer would be that one can always tell whether one is in a given mental state. One cannot. It may be hard to tell whether one's confidence that one knows *p* is high enough for one to count as believing oneself to know *p*, and even harder to tell whether it is rational enough for one to count as rationally believing oneself to know *p* (Chapter 4). There is no good reason to accept a belief-based account of assertion. Indeed, our attitude to false assertions is misrepresented by any simple account on which what warrants assertion does not entail truth.

11.6 MATHEMATICAL ASSERTIONS

The rule of assertion is easier to identify in more formal situations, of which mathematics provides some of the best examples. Assertibility in mathematics has the additional interest that attempts to construct assertibility-conditional theories of meaning have taken the intuitionistic proof-conditional account of mathematics as a paradigm. Assertion in mathematics will therefore be considered. The mathematical case is, it will be argued, more representative than has often been supposed.

In mathematics, the distinction between warranted and unwarranted assertions is striking. Count the propositions that are axiomatic for working mathematicians as having one-line proofs. Then, to a first approximation, in mathematics one has warrant to assert *p* if and only if one has a proof of *p*. On the knowledge account, that is so because, to a first approximation, in mathematics one knows *p* if and only if one has a proof of *p*. One has a proof of *p* when one has followed such a proof and retains some memory of it, in particular of its conclusion. Those are just first approximations, but where having warrant to assert *p* diverges from having a proof of *p*, so does knowing *p*. Conversely, where knowing *p* diverges from having a proof of *p*, so does having warrant to assert *p*. Having warrant to assert *p* and knowing *p* do not diverge from each other; the knowledge account is confirmed.

The word 'proof' has just been used in the informal sense common in ordinary mathematics, in which only truths have proofs; a working mathematician who says that it has been proved that A does not leave it open whether A. This notion is not relativized to an arbitrary formal system; if it were, the connection with (unrelativized) assertibility would be lost. The axioms have one-line proofs in virtue of their status in the practice of mathematics, not in virtue of their place in a particular formal system. 'Proof' will be used in this informal sense below.

Consider first putative cases in which one has warrant to make a mathematical assertion, but lacks a proof. In the simplest cases, one knows by testimony that there is a proof of *p*: but then one knows *p* by testimony, and thereby satisfies the knowledge rule. In rarer cases, the non-deductive evidence for a mathematical proposition may be strong enough to warrant its assertion (Steiner 1975: 93–108). Nevertheless, this biconditional remains plausible: the evidence is strong enough to warrant asserting *p* if and only if it is strong enough for one to know *p*. What the knowledge account will not grant is that one can have warrant to assert *p* without a proof of *p* by having grounds for a mistaken belief that one has a proof of *p*, or for a mistaken belief that there is such a proof. When one has such a belief, on the knowledge account, one at best mistakenly believes that one has warrant to assert *p*. Even if expert mathematicians play a practical joke and inform you falsely that *p* has been proved, you do not really acquire warrant to assert *p* (if you did, the joke would be still less funny).[12] You acquire only misleading evidence that you have such warrant. Although your belief that you have it is reasonable, that does not make it true. The reasonableness in question can be explained as derivative from the knowledge rule. You reasonably believe yourself to know *p*, so you have reason to believe that you have warrant to assert *p*. This view of the matter is independently defensible. Testimony is a special source of warrant because one speaker can *pass on* a warrant to another. Since the expert mathematicians have no warrant to assert *p* themselves, they have none to pass on to you.

Now consider putative cases in which one has a proof of a mathematical proposition but lacks warrant to assert the proposition. The possibility of such cases is sometimes denied, on the Cartesian grounds that genuine proofs are transparent to the subject. That denial does little justice to the complexity of many actual proofs. It can take months of effort by the mathematical community to decide whether a purported proof is genuine. When I have a genuine proof, expert mathematicians may tell me falsely that it contains a fallacy. They may give me a complicated explanation of the supposed fallacy, blinding me with science. I may recall other occasions on which what I believed for broadly similar reasons to be a proof really did turn out to be fallacious. In such cases, it would be unreasonable for me to assert *p*, for it is unreasonable for me to believe that I have warrant to assert *p*. It does not immediately follow that I have no warrant to assert *p*. One may have the authority to do something even when it is unreasonable for one to believe that one

[12] See Kitcher 1983, especially 55–6 and 89–91, for discussion of the social character of mathematical knowledge. Whether this character permits it to be a priori is another matter.

has that authority. What the knowledge account implies is just that I cease to have warrant to assert if and only if I cease to know. That biconditional remains plausible. If I know *p*, I thereby have warrant to assert *p*. Conversely, there is no reason to expect my possession of a proof of *p* to give me warrant to assert *p* independently of letting me know *p*. The more plausible of the two ways for the biconditional to hold is for me to lose both knowledge and warrant to assert: the appearance of ignorance undermines knowledge in a way in which the appearance of knowledge does not undermine ignorance. But even if I retain both knowledge and warrant to assert, the knowledge account stands.

The remaining cases are those in which one has a proof of *p* if and only if one has warrant to assert *p*. They are still less likely to threaten the proposed equivalence between knowing *p* and having warrant to assert *p*.

How untypical are mathematical assertions? Proofs are often supposed to warrant them in a way inapplicable to most or all empirical assertions. Proofs, it is said, are conclusive, whilst empirical warrants are not. However, the nature of the contrast is unclear. No doubt new information cannot make a proof into a non-proof. But the issue is not whether proofs continue to be proofs; it is whether they continue to warrant assertion. Define a way of having warrant to assert *p* to be *defeasible* just in case one can have warrant to assert *p* in that way and then cease to have warrant to assert *p* merely in virtue of gaining new evidence. A way of having warrant to assert *p* is indefeasible just in case it is not defeasible. Most ways of having warrant to make empirical assertions are defeasible, but the considerations above about the social character of mathematical knowledge suggest that even grasping a proof of a mathematical proposition is a defeasible way of having warrant to assert it. One can have warrant to assert a mathematical proposition by grasping a proof of it, and then cease to have warrant to assert it merely in virtue of gaining new evidence about expert mathematicians' utterances, without forgetting anything. If so, mathematical propositions do not differ from empirical ones in point of defeasibility.

The notion of indefeasibility should not be confused with that of factiveness. A way of having warrant to assert *p* is factive just in case a necessary condition of having warrant to assert *p* in that way is that *p* is true. Grasping a proof of a mathematical proposition is a factive way of having warrant to assert it: a necessary condition of grasping a proof of *p* is that *p* is true. Factiveness does not entail indefeasibility. Knowing *p* is always a factive way of having warrant to assert *p*; it is almost never an indefeasible way. New evidence can almost always undermine old knowledge (perhaps there is an indefeasible way of having warrant to assert that one exists). On the knowledge account, any way of having warrant to assert something is factive. Thus mathematical propositions also fail to differ from empirical ones in point of factiveness.[13]

By itself, indefeasibility does not entail factiveness. If warrant to assert *p* consisted merely in good reason to believe *p*, then the inhabitants of a universe created six thousand years ago with every appearance of having existed for millions of years might have an indefeasible non-factive warrant to assert that they are not inhabitants of a universe created six thousand years ago with every appearance of having existed for millions of years. The account defended in this chapter guarantees factiveness independently of indefeasibility.

On the showing of this section, mathematical practice is consonant with the knowledge account in a way that generalizes smoothly to practice outside mathematics.

11.7 THE POINT OF ASSERTION

The knowledge rule is a constitutive rule; it is not a convention. The rule might nevertheless be linked to conventions. Suppose that a language £ assigns to each sentence type *s* in some domain a proposition £(*s*). Then it might be a convention in a particular community that in normal contexts one should utter *s* only if one knows £(*s*). Such a convention of *knowledgeableness* in £ might even be part of what it is for £ to be the language of that community. This convention in £ is an obvious variant on the convention of truthfulness in £ used by David Lewis

[13] The distinction between indefeasibility and factiveness is blurred in this passage from Dummett: 'When we first learn language, we are taught to make assertions only in the most favoured case, namely in that situation in which the speaker can recognize the statement as being true. One way to observe the convention to try to utter only true assertoric sentences would of course be to utter them only in this most favoured situation; but for the great majority of forms of statement that is not at all what we do. Some forms of statement—those in the future tense, for example—are never uttered in the situation which conclusively establishes their truth; and others, which we originally learned to utter only in such situations, we later learn to utter in circumstances in which we may turn out to have been mistaken' (1981: 355). It seems to be assumed that (i) recognizing a truth requires conclusively establishing it and (ii) ordinary standards permit us to assert propositions whose truth has not been conclusively established. If 'conclusively' means indefeasibly, (i) is false. If 'conclusively' means factively, we have no reason to accept (ii). See also McDowell 1982 and Wright 1993 on criteria and defeasibility.

to define what it is for £ to be the language of a community: 'To be truthful in £ is to act in a certain way: to try never to utter any sentences of £ that are not true in £. Thus it is to avoid uttering any sentence of £ unless one believes it to be true in £' (1983: 167, from Lewis 1970). Of course, the account of what it is for £ to be the language of a community must somehow take into account the probability that speech acts other than assertion can be performed in £ (Lewis 1983: 172). The shift from conventions of truthfulness to conventions of knowledgeableness also has repercussions in the methodology of interpretation. The appropriate principle of charity will give high marks to interpretations on which speakers tend to assert what they know, rather than to those on which they tend to assert what is true, or even what is reasonable for them to believe.

It is pointless to ask why the knowledge rule is the rule of assertion. It could not have been otherwise. It is, however, pointful to ask why we have such a speech act as assertion in our repertoire. Could we not have done otherwise? No doubt we need a speech act something like assertion, to communicate beliefs, but could we not have done so just as well by using a speech act whose rule demanded less than knowledge? It would have to permit testimony and inference to enable us to utter new instances on the basis of old ones, just as they do for assertion. But the knowledge rule is not the only rule to underwrite that possibility; the truth rule is another.[14]

One obvious answer is that we need assertion to transmit knowledge.[15] In normal circumstances, when the hearer knows that the speaker asserted *p*, the speaker has no reputation for unreliability, and so on, a speaker who asserts *p* thereby puts a hearer in a position to know *p* if (and only if) the speaker knows *p* (see Lackey 1999 for some qualifications). That answer is probably right, as far as it goes, but leaves at least two points to be explained. First: why could we not transmit knowledge by means of a speech act whose rule required only (for example) truth? The idea might be that, when successful communication occurs, what is transmitted is what is *overt* in the assertion of *p*, and what is overt in the

[14] Brandom 1983 and 1994 give an interesting account of the role of testimony and inference in the social practice of assertion; but see footnote 10. Craig 1990a attempts to explain the concept of knowledge in terms of the need for such a practice, but the knowledge account is not an immediate consequence.

[15] '[T]he essential character of the assertoric use of language lies in its availability for communicating, in the sense of transmitting knowledge about the subject matter of assertions' (McDowell 1980: 128; compare Evans 1982: 310). Recanati 1987: 183 says 'It is a part of our prototype of assertion that if someone asserts that *p*, he knows that *p* and wishes the hearer to share his knowledge'.

assertion is satisfaction of the rule, so what is transmitted is satisfaction of the rule; thus knowledge is transmitted if and only if it is what the rule requires. However, the relevant notion of overtness is hard to pin down. If the overtness of the satisfaction of the rule put the hearer in a position to know that the rule was satisfied, then, if the rule required truth, the hearer would be in a position to know that the assertion was true, and the truth rule would suffice for the transmission of knowledge after all. Second: it will be asked why the transmission of knowledge is what matters, rather than the transmission of true belief, or reasonable belief, or some other cognitive attitude.

A comparison between knowing and doing gives a clue to a further line of thought. One may think of knowing *p* as standing to believing *p* as (intentionally) bringing *p* about stands to desiring *p*. If one knows *p*, then *p* is true; likewise, if one brings *p* about, then *p* is true. But even if *p* was true and one believed *p*, it does not follow that one knew *p*; likewise, even if *p* was true and one desired *p*, it does not follow that one brought *p* about. In each case, the fit between content and world is insufficient because it may have been 'accidental'. Both knowing *p* and bringing *p* about are ways of ensuring *p*; what differs is the direction of fit. If one brings *p* about, one's actions (characterized in environment-dependent ways) ensure the truth of *p*; likewise, if one knows *p*, one's mental states (characterized in environment-dependent ways) ensure the truth of *p* (see section 1.4).

Obedience to a command, as ordinarily understood, involves bringing something about; what matters is not simply the fit between content and world, but someone's responsibility for that fit. To issue a command with appropriate authority is to confer a responsibility; to obey a command is to discharge that responsibility.[16] The point emerges more distinctly for negative commands, where what is commanded is not itself an intentional action. You shout 'Don't move!'; I try to move, but find myself stricken by paralysis. In one sense I did not obey your command. Although its content was fulfilled, I did not ensure that it was; I did not bring it about. The knowledge account extends the analogy between commanding and asserting. To make an assertion is to confer a responsibility (on oneself) for the truth of its content; to satisfy the rule

[16] Obedience to a command may also require an appropriate causal connection between the command and the action; if I do not hear the command to halt, but halt anyway, then I did not intentionally obey the command. Similarly, if I know *p* and assert *p*, but the asserting is causally independent of the knowing, then something is wrong with the asserting. In some sense, in making my assertion I did not fully respect the knowledge rule, although I was lucky enough to get away without violating it.

of assertion, by having the requisite knowledge, is to discharge that responsibility, by epistemically ensuring the truth of the content.[17] Our possession of such speech acts is no more surprising than the fact that we have a use for relations of responsibility.

[17] According to Searle 1969: 66, an assertion of *p* '[c]ounts as an undertaking to the effect that *p* represents an actual state of affairs'. See also Brandom's account (1994: 199–206).

12
Structural Unknowability

12.1 FITCH'S ARGUMENT

The limits of knowledge on which previous chapters concentrated were extrinsic to the propositions unknown. A tree may be at least *n* inches tall although no one is in a position to know that it is at least *n* inches tall, but there is nothing intrinsically unknowable about the proposition that the tree is at least *n* inches tall. If my visual discrimination were better, or I had appropriate measuring instruments, or the tree were taller, then I should be in a position to know that it is at least *n* inches tall. The emphasis was on the virtual ubiquity of such extrinsic limits. They are the rough texture of our cognitive life. This chapter explores some limits to knowledge which are intrinsic to the propositions unknown, necessary limits embedded throughout our contingent ignorance.

Mathematics and physics reveal unexpected limits to our knowledge which depend on Gödelian undecidability, complexity considerations, spatio-temporal limits to the portions of the universe with which we can causally interact, and the like. The limits discussed in this chapter are far more prosaic than those. But they are also more thoroughly intrinsic to the propositions unknown, for they do not depend on our contingent computational limitations or the contingent causal structure of space-time. They arise wherever we are ignorant at all.

The argument for such intrinsic limits was first published by Frederic Fitch (1963), although he attributed it to an anonymous referee in 1945 for a paper which he submitted but never published. The argument was reintroduced into public discussion by Bill Hart and Colin McGinn (1976), whose attention was also drawn to it by an anonymous referee (see also Hart 1979). The nub of the argument is this: if something is an unknown (but perhaps knowable) truth, then that it is an unknown truth is itself an unknowable truth. Every point of contingent ignorance corresponds to a point of necessary ignorance. The core of the argument has already been used at several places in this book without much discussion, for example when section 11.3 deployed Moorean paradoxes in arguing for the knowledge account of assertion.

The argument is sometimes called 'The Paradox of Unknowability', although why it should be regarded as a paradox is quite unclear. The conclusion that there are unknowable truths is an affront to various philosophical theories, but not to common sense. If proponents (and opponents) of those theories long overlooked a simple counterexample, that is an embarrassment, not a paradox. This chapter primarily concerns the positive lessons of Fitch's argument, not its use to refute philosophical theories, although the prospects for such applications will be mooted from time to time.

We must analyse Fitch's argument in more detail. Let *strong verificationism* be the insane-sounding thesis that every truth is known; in symbols:[1]

$$\text{SVER} \quad \forall p(p \supset \mathrm{K}p)$$

Let *weak verificationism* be the sane-sounding thesis that every truth is knowable, in the sense that it is possible for it to be known:

$$\text{WVER} \quad \forall p(p \supset \Diamond \mathrm{K}p)$$

Obviously, since whatever is can be, strong verificationism entails weak verificationism. Fitch's argument shows, on very weak assumptions, that weak verificationism entails strong verificationism.

In the initial presentation of the argument, we will read the operators $\Diamond$ and K as 'it is possible that' and 'it is known that' respectively, without probing their meaning further. Once the argument has been presented, we will ask how we must understand those phrases to make it valid.

To carry through Fitch's argument, we first argue that nothing can be known to be an unknown truth. In brief, if something is known to be an unknown truth then it is known to be a truth; but, equally, if it is known to be an unknown truth then it *is* an unknown truth and therefore is not known to be a truth. The argument uses two principles about knowledge: that it is necessarily factive and that it necessarily distributes over conjunction. Nothing can be known without being true:

$$\text{FACT} \quad \forall p\Box(\mathrm{K}p \supset p)$$

If a conjunction is known, its conjuncts must be known:

$$\text{DIST} \quad \forall p\forall q\Box(\mathrm{K}(p \wedge q) \supset (\mathrm{K}p \wedge \mathrm{K}q))$$

[1] The informal glosses are not strictly synonymous with the formulas themselves, since a truth predicate is a constituent of the former but not the latter. The official argument is given by the formulas; the glosses serve a purely heuristic purpose. Purists will note several departures from accuracy in the informal use of formal notation in this chapter; their purpose is to simplify the exposition.

Substituting $\sim Kp$ for p in FACT gives this special case:

(1) $\forall p\Box(K{\sim}Kp \supset {\sim}Kp)$

Substituting $\sim Kp$ for q in DIST gives this special case:

(2) $\forall p\Box(K(p \wedge {\sim}Kp) \supset (Kp \wedge K{\sim}Kp))$

Since $K{\sim}Kp \supset {\sim}Kp$ is inconsistent with $Kp \wedge K{\sim}Kp$, and the consequences of necessary truths by reasoning in propositional logic are themselves necessary, (1) and (2) yield:

(3) $\forall p\Box{\sim}K(p \wedge {\sim}Kp)$

Since necessary truths are those that are not possibly false, (3) is equivalent to:

(4) $\forall p{\sim}\Diamond K(p \wedge {\sim}Kp)$

We now show that (4) makes weak verificationism collapse into strong verificationism. In brief, if, contrary to strong verificationism, something is an unknown truth, then by (4) it is an unknowable truth that it is an unknown truth, contrary to weak verificationism. More precisely, substituting $p \wedge {\sim}Kp$ for p in WVER gives this special case:

(5) $\forall p((p \wedge {\sim}Kp) \supset \Diamond K(p \wedge {\sim}Kp))$

Given (4), (5) yields:

(6) $\forall p{\sim}(p \wedge {\sim}Kp)$

But (6) is equivalent to SVER by elementary reasoning. Thus 'weak' verificationism is as strong as 'strong' verificationism.

Strong verificationism is obviously false. Of course, (4) itself implies that we cannot know of any one proposition that it is a counterexample to strong verificationism, for then we should be in a position to know it to be an unknown truth, which by (4) we cannot do. But we can do the next best thing: we can know of two propositions that one or other of them is an unknown truth; we just cannot know which. For example, either my office contains an even number of books at noon on 11 October 1999 (time *t*) or it does not. I could find out by counting whether it contains an even number of books at *t*. But I will not count them; nor will anyone else. As a matter of contingent fact, no one will ever know whether my office contains an even number of books at *t*. Thus either it is an unknown truth that my office contains an even number of books at *t* or it is an unknown truth that my office contains an odd number of books at *t*. Either way, there is an unknown truth; strong verificationism is false. Fitch's argument then shows that either it is an unknowable truth that it is an unknown truth that my office contains an even number of books at *t* or it is an unknowable truth that it is an unknown truth that my office contains an odd number of books at *t*. Either way, there is an unknowable truth; weak verificationism is false too. Although strong verificationism is obviously false and weak verificationism is not obviously false, the two theses are equivalent, given our assumptions.

We can construct similar counterexamples to strong verificationism from Gray's stanza:

Full many a gem of purest ray serene,
  The dark unfathom'd caves of ocean bear:
Full many a flower is born to blush unseen,
  And waste its sweetness on the desert air.

Each of them corresponds to a counterexample to weak verificationism. Quite generally, each counterexample to strong verificationism corresponds to a counterexample to weak verificationism.

How should we understand the operators $\Diamond$ and K in Fitch's argument? The weaker the proposition $\Diamond Kp$, the weaker the thesis WVER, so the more significant is its refutation. Since the strength of $\Diamond Kp$ increases monotonically with the strength of Kp, we seek the weakest interpretations of $\Diamond$ and K that validate the argument.

For the step from (3) to (4), $\Diamond$ should express a kind of possibility that is dual to the kind of necessity expressed by $\Box$; $\Diamond$ is equivalent to $\sim\Box\sim$ and $\Box$ to $\sim\Diamond\sim$. For FACT and DIST, knowledge should be factive and distribute over conjunction with that kind of necessity. Moreover, for the step from (1) and (2) to (3), that kind of necessity should be preserved by deductions in classical propositional logic. These requirements constrain the relevant kind of necessity hardly at all. Perhaps it is not a purely logical necessity that knowledge is factive and distributes over conjunction, since the notion of knowledge is not a purely logical one. But it might still have those features as a matter of metaphysical necessity: however things had been, knowledge would still have been factive and distributed over conjunction. We will therefore understand $\Diamond$ and $\Box$ as 'it is metaphysically possible that' and 'it is metaphysically necessary that' respectively.

We can understand K as 'some being at some time [past, present or future] knows that'. Since those who believe that there actually is an infinite omniscient being will take SVER to be true on that understanding, they should understand 'being' as tacitly qualified by 'finite'. We must not allow the phrase 'at some time' to act as a tense operator on the content clause following 'that'; it merely binds the time variable in 'knows'. Thus K(it is raining) is evaluated as true with respect to a time *t* if some being knows at some time t^* of rain at *t*; knowledge at t^* of rain at t^* is neither necessary nor sufficient. This is in fact a natural reading of the English sentence 'It was, is, or will be known that it is raining'. On the other reading, K(it is raining) would entail 'It was, is, or will be raining' but not 'It is raining' itself; thus FACT would fail.

On these understandings, Fitch's argument concludes that for some truth *p*, however things had been, no being would ever have known *p*. We can also substitute other attitudes for knowledge in the interpretation of K. Any necessarily factive attitude that necessarily distributes over conjunction will do, although we must check that SVER is genuinely implausible on the new interpretation.

Consider, for example, *T-conception.* One T-conceives *p* if and only if *p* is true and one conceives (that is, grasps the proposition) *p*. T-conception is factive by definition. It necessarily distributes over conjunction because both its conjuncts do; a true conjunction has true conjuncts, and in conceiving the conjunction one conceives the conjuncts. Fitch's argument therefore shows that if all truths are T-conceivable, then all truths are T-conceived. But surely not all truths are T-conceived, for not all truths are conceived; some will never be grasped by anyone. Naturally, one cannot conceive a specific example of a truth which no one will ever conceive. One cannot even do what can be done in the case of knowledge, by listing two or more propositions such that one knows that at least one item on the list is an example. For in listing the propositions in the intended sense, one would conceive all of them. Nevertheless, there are examples of never conceived truths, and we can gesture towards areas in which some of them lie. For example, there are conceivable truths which no one will ever conceive about the state of my office at noon on 11 October 1999. We all have better things to think about. By contraposition, the argument shows that there are truths about the state of my office at that time which can be conceived, but can never be both true and conceived.

Can the argument be generalized to non-factive attitudes? It uses FACT only to derive ~K*p* from K~K*p*, and could therefore make do with that restricted subcase of factiveness. For example, if one reads K as 'it is rational to believe that', one might hold that although it is sometimes rational to believe a false proposition, if it is rational to believe that it is not rational to believe *p* then it is not rational to believe *p*. If what it is rational to believe is closed under conjunction, a counterexample to that principle would involve its being rational to believe a Moore-paradoxical conjunction (*p* and it is not rational to believe *p*). But such principles are highly sensitive to the interpretation of 'rational to believe'. For instance, an example in section 10.5 shows that even if p is at least 80 per cent probable on one's evidence, it may still be at least 80 per cent probable on one's evidence that p is not at least 80 per cent probable on one's evidence. Perhaps the required inference goes through on other interpretations of 'rational to believe', on which the argument shows that some truths cannot be rationally believed. Since our concern is primarily with knowledge, we can leave that question open.[2]

Fitch's argument against weak verificationism employs classical logic. Dummett's well-known arguments for weak verificationism ('anti-realism') commit the weak verificationist to rejecting classical logic in favour of intuitionistic logic, or something like it. Fitch's argument is not intuitionistically valid. In particular, the inference from (6) to SVER involves the intuitionistically invalid move from $\sim(p \wedge \sim Kp)$ to $p \supset Kp$. Thus Dummett's weak verificationist is not committed to strong verificationism, although the commitment to (6) may reasonably be regarded as already bad enough. At any rate, the use of Fitch's argument against Dummett's weak verificationist is not dialectically straightforward. Fortunately, section 4.8 identified a flaw in Dummett's reasoning (there are probably others), and Dummettian anti-realism is in any case not the main topic of this book.[3] For present purposes, we assume without argument the validity of classical logic. Our aim here is not to outmanœuvre an anti-realist opponent but to explore limits to knowledge from a classical starting point. The next two sections confront objections to Fitch's argument within a classical framework. The final sections discuss modifications of weak verificationism designed to finesse the argument.

12.2 DISTRIBUTION OVER CONJUNCTION

Fitch's argument assumes that knowing the conjuncts is necessary for knowing a conjunction (DIST). Once DIST is dropped, the remaining background assumptions do not permit the reductio ad absurdum of weak verificationism, the derivation of SVER from FACT and WVER.

We can clarify the role of DIST in Fitch's argument by giving K an

[2] For further discussion of the applications of Fitch-type arguments see Daniels 1988, Edgington 1985, Kvanvig 1995, MacIntosh 1984, Mackie 1980, Rescher 1984, Routley 1981, Sorensen 1988, and references therein. For a related argument see Plantinga 1982.

[3] For discussion of Fitch's argument in the context of intuitionistic logic see Cozzo 1994, Pagin 1994, Percival 1990, Tennant 1997: 245–79, Usberti 1995: 65–6, 121–8, Williamson 1982, 1988, 1992c, 1994a, 2000e, Wright 1993: 427–30, and references therein.

unintended reinterpretation. We introduce a new sentence constant c into the language and consider possible worlds models in which, for every formula A, KA is interpreted as $A \wedge (c \vee \sim\Diamond(A \wedge c))$. All the other operators are treated as usual. In particular, $\Diamond$A is true at a world w if and only if A is true at some world to which w bears the accessibility relation in the model. FACT is trivially true at all worlds in all such models, for KA is interpreted as though it had A as a conjunct. WVER is also true at all worlds in all such models, provided that $\Diamond$A is true at a world whenever A is, which we can guarantee by making accessibility reflexive. For suppose that p is true at a world w. There are just two cases to consider. If $\Diamond(p \wedge c)$ is true at w, then $p \wedge c$ is true at some world x accessible from w, so $p \wedge (c \vee \sim\Diamond(p \wedge c))$ is true at x, so Kp is interpreted as true at x, so $\Diamond$Kp is interpreted as true at w. If $\Diamond(p \wedge c)$ is false at w, then $p \wedge (c \vee \sim\Diamond(p \wedge c))$ is true at w, so Kp is interpreted as true at w, so $\Diamond$Kp is interpreted as true at w, since accessibility is reflexive. Either way, $p \supset \Diamond$Kp is interpreted as true at w, as WVER requires. But we can easily construct such models in which SVER fails. For example, if p is necessarily true at a world w and c is contingently false at w, then Kp is interpreted as false at w. DIST fails because K$(p \wedge \sim c)$ is interpreted as true at w. Any reasonable modal logic for the interpretation of $\Diamond$ and $\Box$ as metaphysical possibility and necessity respectively has such models. In particular, there are such models of the strong modal logic S5 augmented with propositional quantifiers like those used in the argument (see Fine 1970 and Kaplan 1970). Thus SVER cannot be derived even from the necessitation of WVER and FACT in propositionally quantified S5, let alone the vastly weaker modal logics that suffice for Fitch's argument. Of course, we can easily point to differences between the deviant interpretation of K and the intended one. The point is simply that FACT by itself provides insufficient information about K to enable us to make the connection from WVER to SVER.[4] Something more is needed, such as DIST.

Is there any reason to doubt the distribution principle DIST? Many propositional attitudes distribute over conjunction. In conceiving a conjunction, for example, one conceives its conjuncts; in asserting a conjunction, one presumably asserts its conjuncts. But not all attitudes distribute. I disbelieve the conjunction that Edinburgh is in Scotland and Scotland is in Asia because I disbelieve the latter conjunct, but I do not disbelieve the former. However, that example leaves open the possi-

[4] Williamson 1990b proves in effect that SVER is derivable from the necessitations of WVER and its converse in the absence of DIST given S4 as the background modal logic but not given the weaker background modal logic KT (although the systems considered there lack propositional quantification).

bility that all *positive* propositional attitudes distribute over conjunction, in a sense of 'positive' that would need to be made precise.[5]

If a positive propositional attitude is closed under at least some forms of logical consequence beyond logical equivalence, we may expect it to be closed under a very intimate one such as the $\wedge$-elimination inference from $p \wedge q$ to p and to q. To investigate whether it distributes over conjunction would then be to test the null hypothesis that it satisfies no form of deductive closure at all, or at least none involving inferences between logically inequivalent formulas. The null hypothesis in turn exemplifies a liberal view of the extent to which rationality constrains the attribution of propositional attitudes (of course, it is not suggested that closure under logical consequence is even a rational ideal for negative propositional attitudes, such as disbelief). Our present interest is in a specific positive propositional attitude, knowing.

Robert Stalnaker has shown that there is much more to be said than we might have expected for the claim that both knowledge and belief are closed under logical consequence (1984: 71–99 and 1999: 241–73). If the claim were correct, a fortiori DIST would hold. But if DIST depends on the deductive closure of knowledge, its status is shaky indeed, for Stalnaker's highly qualified considerations do not destroy the plausibility of the original case against closure. Let us retain the common-sense idea that, in logic and mathematics, we may understand a conjecture p without knowing p until we prove p, even though all along it was in fact a logical consequence of what we knew. If we do not know all the logical consequences of what we know, we do not even know all the *obvious* logical consequences of what we know, for what we know is linked to some of what we do not by chains of propositions each member of which is an obvious logical consequence of its predecessors.[6] Thus to deny DIST is not to deny that a conjunction obviously entails its conjuncts. Our question is whether a conjunction has some more intimate relation to its conjuncts under which knowledge is closed, even though it is not closed under obvious logical consequence.

As usual, we assume that knowledge entails belief. Then two quite different kinds of putative counterexample might be offered against the claim that knowledge distributes over conjunction. Suppose that one

[5] The generalization is suggested by Barwise and Perry 1983: 205.

[6] Contrast Rescher 1968: 46–7, who suggests that belief is closed under obvious consequence, and therefore distributes over conjunction; since the obvious consequence relation is not transitive, he requires the principle to be used at most once in an argument. But that makes the consequence relation itself non-transitive. If some arguments are truth-preserving, so is the result of chaining them together. See also Hintikka 1975. In the case of belief, Brian Loar proposes distribution over conjunction as a perhaps dubious and approximate constraint on a functionalist attribution of beliefs (1981: 72).

knows $p \wedge q$ without knowing p. Then one believes $p \wedge q$. One either believes p or fails to believe p. If one fails to believe p, that by itself accounts for one's failure to know p. Call such a case *irrational*. It would also be a counterexample to the principle that belief distributes over conjunction. On the other hand, if one believes p, then one's belief is true, since $p \wedge q$ is true; one's failure to know p would have to be accounted for in some other way. Call such a case *rational*. With respect to the relevant propositions it would not violate the distribution of belief over conjunction.

Formally, the distribution of belief over conjunction is neither necessary nor sufficient for the distribution of knowledge over conjunction. It is unnecessary, because belief in the conjunct might somehow follow from knowledge of the conjunction, although not from mere belief in it; perhaps failure to note certain immediate consequences blocks knowledge but not belief. Thus knowledge might distribute while belief failed to do so. Conversely, the distribution of belief is insufficient for the distribution of knowledge, since it rules out irrational counterexamples to the latter principle but not rational ones. Are there rational or irrational counterexamples?

Robert Nozick's analysis of knowledge implies the possibility of rational counterexamples to the distribution of knowledge over conjunction. It does not imply the possibility of counterexamples to the distribution of belief. According to Nozick, knowledge is truth-tracking belief. On the simple version of his analysis, before methods complicate matters, one knows p if and only if (1) p is true; (2) one believes p; (3) if p were false one would not believe p; (4) if p were true but things were slightly different one would still believe p. On this analysis, belief in a conjunction may track the truth while belief in a conjunct does not because '[w]e can satisfy condition 3 for a conjunction by satisfying it for its most vulnerable conjunct, the one that would be false if the conjunction were false; it does not follow that we satisfy condition 3 for the other conjunct as well' (1981: 228).[7] Nozick sometimes knows that he is in Emerson Hall and not floating in an Emerson-Hall-simulator on

[7] Nozick 1981: 692 n.63 argues that the distribution of knowledge over conjunction would enable one to derive (i), the closure of knowledge under known logical implication, from (ii), the closure of knowledge under known logical equivalence. Since he accepts (ii) and rejects (i), that would give him another motive for denying distribution. However, his argument relies on the assumption that whenever he knows that r entails s, he knows that r is logically equivalent to $s \wedge (s \supset r)$. That is only as plausible as the assumption that whenever he knows that r is logically equivalent to s, he knows that r is logically equivalent to $s \wedge (s \supset r)$. But, by iteration, that assumption entails that whenever he knows that r is logically equivalent to s, he knows that r is logically equivalent to $(s \wedge (s \supset r)) \wedge ((s \wedge (s \supset r)) \supset r)$, and so on ad infinitum, which is implausible.

Alpha Centauri, for if the conjunction had been false it would have been so because he was somewhere mundane other than Emerson Hall, not because he was floating in an Emerson-Hall-simulator on Alpha Centauri, and he would then not have believed the conjunction. However, at those times he does not know that he is not floating in an Emerson-Hall-simulator on Alpha Centauri, for if that conjunct had been false he would still have believed it. All of this is quite consistent with the distribution of belief over conjunction—although if one believes that one knows $p \wedge q$, that one does not know p and that one should believe only what one knows, one may try to believe $p \wedge q$ without believing p.

We saw reason in Chapter 7 to reject Nozick's account of knowledge, even when it is complicated by reference to methods. Nevertheless, we can consider a putative counterexample of Nozick's kind to an instance of DIST of the sort used in Fitch's argument, to decide whether the case has any independent plausibility. A climber reaches the summit of a mountain at 12.03, but does not look at his watch until much later; nobody else is around. Let p be the proposition that he did not reach the summit between 12.01 and 12.02. He is convinced of p on probabilistic grounds, but neither he nor anyone else will ever know p. If p had been false, he would still have believed p on the same probabilistic grounds. Indeed, he knows that no one will ever know p. He truly believes the conjunction $p \wedge {\sim}Kp$, and his belief tracks the truth, for ${\sim}Kp$ is the more vulnerable conjunct. He could have looked at his watch more easily than he could have reached the summit earlier. If $p \wedge {\sim}Kp$ had been false, ${\sim}Kp$ would have been false, so Kp would have been true; he would then have believed Kp and would not have believed $p \wedge {\sim}Kp$; thus condition 3 is satisfied. Moreover, if $p \wedge {\sim}Kp$ had been true while things were slightly different, he would still have believed $p \wedge {\sim}Kp$; thus condition 4 is satisfied. On Nozick's account, $K(p \wedge {\sim}Kp)$ is true and Kp false. But this is not a convincing example of knowing a conjunction without knowing the conjuncts, for, intuitively, the merely probabilistic grounds on which the climber believes p prevent him from knowing $p \wedge {\sim}Kp$ as well as from knowing p. The intuition is general: in such cases one has inadequate grounds for one's true belief in a conjunction only if one has inadequate grounds for one's true belief in at least one of the conjuncts. The conflict between Nozick's analysis of knowledge and the distribution principle is a problem for Nozick's analysis, not for distribution.

What of irrational counterexamples to distribution? We can easily find cases in which someone would assent to a conjunction if queried but would have dissented from (and not assented to) one conjunct if queried on that alone. For example, someone might answer 'Yes' if asked 'Is it true that a city other than Rome was once the capital of Italy and Turin was the capital of Italy from 1861 to 1864?' although he would have answered 'No' if asked 'Is it true that a city other than Rome was once the capital of Italy?'. He might have forgotten that Turin was the capital of Italy from 1861 to 1864, and need mention of it to jog his memory. If a disposition to heartfelt assent on being asked whether *p* is true is sufficient for believing *p*, and a disposition to heartfelt dissent in those circumstances is incompatible with believing *p*, then our man believes that a city other than Rome was once the capital of Italy and Turin was the capital of Italy from 1861 to 1864 although he does not believe that a city other than Rome was once the capital of Italy. Moreover, he knows the conjunction, having learnt it at school from a reliable teacher in the usual way. Yet he does not know the conjunct, for he does not believe it. But the case is not terribly convincing, for one's speech dispositions are an inadequate test of belief. Arguably, our man does have a non-occurrent belief that a city other than Rome was once the capital of Italy, which sometimes slips his conscious mind. As for occurrent belief, when he occurrently believes the conjunction, he occurrently believes the conjuncts.

A subtle challenge to distribution arises indirectly from analogies between the semantic paradoxes and some epistemic paradoxes (Kaplan and Montague 1960, Burge 1978 and 1984, Koons 1992). Here is an example of such a paradox. I say at time *t* just 'I am not expressing knowledge at *t*'. If I am expressing knowledge at *t*, then I am expressing knowledge that I am not expressing knowledge at *t*; since knowledge is factive, I am therefore not expressing knowledge at *t*. Thus, given the circumstances, the supposition that I am expressing knowledge at *t* is self-defeating. Therefore, I am not expressing knowledge at *t*. Since I am aware at *t* of this reasoning, I know at *t* that I am not expressing knowledge at *t*, so in saying 'I am not expressing knowledge at *t*' at *t*, I express knowledge at *t* after all. That is a contradiction. Such reasoning obviously bears a close resemblance to the Liar Paradox. Many diagnoses of the fallacy have been offered. Suppose that one is sympathetic to a hierarchical solution to the Liar, on which each level *i* of a hierarchy corresponds to a truth predicate 'true$_i$', where something is true$_i$ only if it contains no occurrence of 'true$_j$' for any level *j* not lower than *i* in the hierarchy, and legitimate occurrences of the unsubscripted 'true' in ordinary language are interpreted as implicitly indexed to a level in the hierarchy. Since such an approach can be shown to block the paradoxical reasoning in the Liar, one might adopt a similar approach to the epistemic paradoxes. Legitimate occurrences of the unsubscripted 'know' in ordinary language would be interpreted as implicitly indexed to a level in the hierarchy. I could then be coherently described as knowing$_{i+1}$ (but not knowing$_i$) at *t* that I am not expressing knowledge$_i$ at *t*. We may assume that for each level *i*, knowledge$_i$ is factive and distributes over conjunction; FACT and DIST hold when K in them is subscripted uniformly. We may also assume that knowledge$_i$ entails knowledge$_j$ whenever $i \leq j$; the hierarchy is cumulative. The weak verificationist principle is read as a schema $A \supset \Diamond KA$, in any instance of which K is assigned the least subscript higher than any subscript in A. The result in the special case needed for Fitch's argument would be of the form $(p \wedge \sim K_i p) \supset \Diamond K_{i+1}(p \wedge \sim K_i p)$. By FACT and DIST for K_{i+1} we can derive $(p \wedge \sim K_i p) \supset \Diamond(K_{i+1}p \wedge \sim K_i p)$, but since $K_{i+1}p \wedge \sim K_i p$ is not a contradiction, we cannot proceed to deny $p \wedge \sim K_i p$. Thus an instance $K(A \wedge B) \supset (KA \wedge KB)$ of the original DIST can fail when we assign each occurrence of K the least subscript higher than any subscript in the sentence to which it is applied, since K will be assigned a lower subscript in KA than in $K(A \wedge B)$ if there are higher subscripts in B than in A. Exactly that happens when B is ~KA, as in Fitch's argument.

A difficulty emerges for the hierarchical approach when we ask how $K_{i+1}p \wedge \sim K_i p$ could be true. The answer would be easy if the subscript *i* occurred in *p*, for then K would need a higher subscript than *i* to be correctly applied to *p*. But in the crucial cases, no such subscript occurs in *p*. For example, *p* may say that the number of books in my office at noon on 11 October is even. In what sense could that be known at level *i*+1 but not at level *i*? Perhaps a claim could be known at level *i*+1 but not at level *i* if the route to knowing it involved claims about knowledge$_i$, even though the target claim did not, but it would be bizarre if such contrived cases were crucial to a defence of weak verificationism. They would at any rate not serve a defence based on the Dummettian meaning-theoretic idea that truth implies the possibility of *canonical* verification, for one canonically verifies a conjunction by canonically verifying its conjuncts, and canonically verifying that the number of books in my office at noon on 11 October is even would consist in something like counting them, not in verifying the steps of an indirect argument involving claims about knowledge$_i$. Thus $K_{i+1}p \wedge \sim K_i p$ is an implausible combination.

The hierarchical objection to Fitch's argument faces another problem. We seem able to grasp the idea that *p* is *totally unknown*, in a sense which entails that *p* is unknown$_i$ for each level *i*, but which does not entail that *p* is untrue. If so, we can simply adapt Fitch's argument by considering the proposition that *p* is a totally unknown truth, since that proposition cannot be known$_i$ for any level *i*. Naturally, such quantification over levels must be handled with great care, but we should not allow our precautions to blind us to the ready intelligibility of the supposition that it is a totally unknown truth that the number of books in my office at noon on 11 October 1999 is even. Although it is far from obvious how to solve the epistemic paradoxes, it is most unlikely that an adequate solution would involve restrictions draconian enough to block every form of Fitch's argument.[8]

Can we argue that knowledge *does* distribute over conjunction? It would scarcely be relevant to argue that one will always infer the conjuncts from the conjunction, for, since making an inference takes time, one might still violate distributivity before completing the inference. If knowledge of the conjunction causes knowledge of the conjuncts, and the cause is not simultaneous with the effect, then for an intermediate period one would know the conjunction without knowing the conjuncts; if no one else had the relevant knowledge, that period would be a counterexample to distribution. The members of a rather dim community may take several seconds to notice the entailment from a given conjunction to one of its conjuncts. Moreover, there is no form of inference that one can be relied on to carry out exceptionlessly. Distraction or sudden death is always liable to intervene. The required premise is precisely not that deductive inference is a way of *extending* knowledge. Rather, what would need to be shown is that knowledge of a conjunction is *already* knowledge of its conjuncts.

It might be suggested that one can know a conjunction only by inferring it from its conjuncts, and that the output of the inference is knowledge only if the inputs are, so that in order to know a conjunction one must already know its conjuncts. Perhaps inference from the premises p and q is in some sense the canonical way of coming to know the conjunction $p \wedge q$, since the meaning of $\wedge$ is so closely tied to the validity of the $\wedge$-elimination rule. Nevertheless, there are other ways of knowing $p \wedge q$: from testimony to the conjunction, or by inference from the premises $(p \wedge q) \vee r$ and $\sim r$, and so on.

Even so, $\wedge$-elimination has a rather special status. It may be brought out by a comparison with the equally canonical $\vee$-introduction inference to the disjunction $p \vee q$ from the disjunct p or from the disjunct q.

[8] See Williamson 1990b: 311–12 for further issues about the hierarchical approach. Zemach 1987 has a different objection to DIST, but it depends on what he calls a *de re* reading of Kp on which it requires only that the proposition p be known to be true, possibly by someone who refers to the proposition indirectly without knowing what its content is; in that sense, one can know $p \wedge q$ to be true without knowing that it is a conjunction. Zemach rejects the more natural *de dicto* reading of Kp on the basis of an argument that seems to trade on an equivocation in 'knowing the content' of p, between knowing what the content is and having knowledge of which it is the content (1987: 530).

Although the validity of $\vee$-introduction is closely tied to the meaning of $\vee$, a perfect logician who knows *p* may lack the empirical concepts to grasp (understand) the other disjunct *q*. Since knowing a proposition involves grasping it, and grasping a complex proposition involves grasping its constituents, such a logician is in no position to grasp $p \vee q$, and therefore does not know $p \vee q$. In contrast, those who know a conjunction grasp its conjuncts, for they grasp the conjunction.[9] Moreover, they grasp the sense of its composition: they grasp the conjuncts as consequences of the conjunction. If they know the conjunction they grasp it, and in doing so grasp its conjuncts as its consequences. There is no obstacle here to the idea that knowing a conjunction *constitutes* knowing its conjuncts, just as, in mathematics, we may *count* a proof of a conjunction as a proof of its conjuncts, so that if $p \wedge q$ is proved then *p* is proved, not just provable.

Someone might know a conjunction but believe one of its conjuncts for independent and bad reasons. The latter belief would not constitute knowledge. Nevertheless, the case is no counterexample to what has just been said. For the person may be considered as having two beliefs in the conjunct, one constituted by the knowledge of (and therefore belief in) the conjunction, the other dependent on the bad reasons. The former counts as knowledge; the latter does not.

We have no well-confirmed analysis of knowledge to use as an effective test of the claim that knowledge distributes over conjunction. Indeed, according to the account developed in Chapter 1, knowledge has no analysis of the traditional kind. That account is consistent with the distribution principle but does not entail it. Although DIST is highly plausible, and no objection to it has proved persuasive, the case for it is not quite as decisive as we might hope. It is therefore prudent to ask whether we can modify Fitch's argument so as to avoid commitment to distribution. Since SVER is not derivable from WVER and FACT without DIST in any reasonable modal logic (such as propositionally quantified S5), we must either strengthen the premises or weaken the conclusion. Both strategies look promising.

We can strengthen weak verificationism to moderately weak verificationism, defined as the principle that a conjunction is true only if it is possible that both conjuncts are known:

$$\text{MWVER} \qquad \forall p \forall q((p \wedge q) \supset \Diamond(\mathrm{K}p \wedge \mathrm{K}q))$$

[9] The endorsement by Dretske (1970: 1009) of the claim that knowledge of a conjunction entails knowledge of the conjuncts is compromised by his bracketing of it with the claim that knowledge of a disjunct entails knowledge of the disjunction.

Given a true conjunction, weak verificationism entails that both conjuncts are possibly known ($\Diamond Kp \wedge \Diamond Kq$); moderately weak verificationism adds the compossibility of those possibilities. It is hard to guess what argument someone might have for weak verificationism that would not also be an argument for moderately weak verificationism. For example, Dummett and others motivate WVER by a theory of meaning in which the key concept is canonical verification in place of truth. Rather than saying that $p \wedge q$ is true if and only if p is true and q is true, the theory will say that $p \wedge q$ is canonically verified if and only if p is canonically verified and q is canonically verified. Such a verificationist will argue for WVER by arguing on meaning-theoretic grounds that truth entails the possibility of canonical verification, conceived as a species of knowledge; but then, since canonical verification distributes over conjunction, the truth of a conjunction entails the possibility of canonically verifying every conjunct together, as in MWVER. Now as a special case of MWVER we have $\forall p((p \wedge {\sim}Kp) \supset \Diamond(Kp \wedge K{\sim}Kp))$; but FACT and the background modal logic yield $\forall p{\sim}\Diamond(Kp \wedge K{\sim}Kp)$, so we can derive SVER as before. Thus moderately weak verificationism entails absurdly strong verificationism even in the absence of DIST. Moderately weak verificationism is false. A conjunction may be true even if its conjuncts cannot be known together.

Alternatively, we can weaken strong verificationism to moderately strong verificationism, defined as the principle that a proposition is true only if it is a conjunct of a known conjunction:

MSVER $\forall p(p \supset \exists qK(p \wedge q))$

We can derive moderately strong verificationism from the original weak verificationism even in the absence of DIST. For if p is defined as *completely unknown* if and only if it is not a conjunct of any known conjunction ($p \wedge {\sim}\exists qK(p \wedge q)$), then WVER entails that p is a completely unknown truth only if it can be known that p is a completely unknown truth. But nothing can be known to be a completely unknown truth ($\forall p{\sim}\Diamond K(p \wedge {\sim}\exists qK(p \wedge q))$), for, necessarily, something is known to be a completely unknown truth only if it is not completely unknown. Therefore, given WVER, no truth is completely unknown, a principle equivalent to MSVER. The formal version of this reasoning uses FACT and a background modal logic but not DIST. Now moderately strong verificationism is almost as absurd as strong verificationism. For example, no one will ever know a conjunction of which it is a conjunct that the number of books in my office at noon on 11 October 1999 is even, and no one will ever know a conjunction of which it is a conjunct that the number of books in my office then is odd. Since one of those propositions about the number of books is true, one of them is a completely unknown truth. Thus we can argue without appeal to DIST that even the original weak verificationism is false.

Dorothy Edgington (personal communication) has pointed out another way of modifying Fitch's argument to avoid commitment to DIST. His argument still goes through if we weaken DIST to its consequence $\forall p \forall q(\Diamond \mathrm{K}(p \wedge q) \supset \Diamond(\mathrm{K}p \wedge \mathrm{K}q))$. But what justifies the latter if not DIST? If the idea is that someone who knows the conjunction could have inferred the conjuncts, its natural expression is at most $\forall p \forall q \Box(\mathrm{K}(p \wedge q) \supset \Diamond(\mathrm{K}p \wedge \mathrm{K}q))$, which yields $\forall p \forall q(\Diamond \mathrm{K}(p \wedge q) \supset \Diamond\Diamond(\mathrm{K}p \wedge \mathrm{K}q))$ and its special case $\forall p(\Diamond \mathrm{K}(p \wedge {\sim}\mathrm{K}p) \supset \Diamond\Diamond(\mathrm{K}p \wedge \mathrm{K}{\sim}\mathrm{K}p))$ given just the weak principles of modal logic used in the original argument: that necessity is closed under logical consequence and that possibility is the dual of necessity. If we slightly strengthen FACT by substituting $\Box\Box$ for $\Box$ we can derive $\forall p {\sim}\Diamond\Diamond(\mathrm{K}p \wedge \mathrm{K}{\sim}\mathrm{K}p)$ using the same principles of modal logic, and the argument then proceeds as before to collapse WVER into SVER. Although Edgington's weakening of DIST deals plausibly with irrational counterexamples to distribution in which one accidentally fails to notice the consequences of one's knowledge, it scarcely addresses rational counterexamples, in which one has already inferred the conjuncts but fails to know them for some other reason. The earlier modifications of Fitch's argument respond more pertinently to such cases.

Appendix 6 discusses the earlier modifications of Fitch's argument in a more formal context. For present purposes, the upshot is that rejecting the distribution principle leaves the overall philosophical position unchanged. It is no way out for the verificationist.

12.3 QUANTIFICATION INTO SENTENCE POSITION

As formulated in section 12.1, Fitch's argument involves quantification into sentence position with the propositional variables p and q. When (1) is derived from FACT and (2) from DIST, the complex formula ${\sim}\mathrm{K}p$ is substituted for the universally quantified propositional variables within the scope of the modal operator $\Box$. Such substitutions may appear problematic.

Consider a putative analogy: the inference from 'For every natural number *n*, it is not necessary that *n* is the number of the planets' to 'It is not necessary that the number of the planets is the number of the planets', which apparently involves the substitution of the definite description 'the number of the planets' for the universally quantified variable *n*. Let us assign all the definite descriptions narrow scope, so that the premise is true if and only if for every natural number *n*, in some possible world there are not exactly *n* planets. Thus the premise is indeed true, for it is contingent how many planets there are. But the conclusion is false, for in every possible world there are exactly as many planets as there are planets. Consequently, the inference is invalid. A common diagnosis is that the definite description 'the number of the planets' is not a rigid designator; it designates different numbers with respect to different possible worlds. The moral is then drawn that the rule of universal instantiation does not permit the substitution of non-rigid designators for variables within the scope of modal operators. Since variables are interpreted rigidly, whatever is substituted for them must also be interpreted rigidly, if universal instantiation is to be valid.

We might expect a corresponding restriction to apply to the rule of universal instantiation for quantifiers into sentence position. Then $\sim Kp$ could be substituted for the propositional variables in Fitch's argument only if $\sim Kp$ counted as a rigid designator in whatever sense is appropriate to sentences. No check was made to see that $\sim Kp$ satisfied such a restriction. This has been regarded as a crucial flaw in Fitch's argument (Kvanvig 1995).

To make sense of the question whether $\sim Kp$ is a rigid designator, we must understand what sentences designate with respect to worlds. A suggestion inspired by Frege is that a sentence designates its truth-value at a world with respect to that world. But then a sentence counts as a rigid designator only if it has the same truth-value with respect to every world, which drastically restricts the class of rigidly designating sentences. It certainly excludes $\sim Kp$ in the cases of interest, since it is supposed to be contingent whether *p* is known; for example, it is contingent whether it is known that the number of books in my office at time *t* is even. But that interpretation is far too restrictive. Since the restriction on universal instantiation is needed only if propositional variables are rigid designators, and the objection assumes that the restriction is needed, it presumably counts the formula $\forall p(\Box p \vee \Box\sim p)$ as valid. Propositional quantification into modal contexts is not worth having at that price. Moreover, we surely have an understanding of the formula $\forall p(\Box p \vee \Box\sim p)$ on which it is falsified by contingency.

A more promising suggestion is to treat sentences as designating propositions with respect to worlds. Thus a sentence is a rigid designator if it designates the same proposition with respect to every world, even if that proposition varies in truth-value from world to world. The question is now: why should one doubt that a sentence is rigid? In particular, ~K*p* seems to be as rigid as the propositional variable *p*; with respect to each world it designates the proposition that the proposition *p* is not known.

The formula ~K*p* denies that some being at some time knows *p*. Someone might think that the designation of the quantifiers 'some being' and 'some time' with respect to a world depends on what beings and times there are in that world; if the designation of the quantifier is a component of the proposition designated by the whole formula, and it is contingent what beings or times there are, then the formula will designate different propositions with respect to different worlds. Kvanvig (1995) regards such non-rigidity as a kind of indexicality, like the referential context-dependence of 'I', but that is a confusion. In the terminology of Kaplan (1989), indexicality is variation in reference (designation) with respect to the context in which the expression is uttered, whereas non-rigidity is variation in reference with respect to the circumstance with respect to which it is evaluated. The indexicals 'I' and 'me' are rigid designators; when uttered by a given speaker, they rigidly designate that speaker. For example, although 'James Mill fathered me' was true as uttered by John Stuart Mill, that is irrelevant to the truth-value of 'James Mill could have fathered me' as uttered by Harriet Taylor. In evaluating her supposed utterance, we take 'me' rigidly to designate her, and ask whether in any (possible) circumstances James Mill fathered her. We do not consider contexts in which 'me' designates someone other than Harriet Taylor. Similarly, contextual variation in the reference of ~K*p* is irrelevant to present concerns. What matters is whether ~K*p*, as uttered in a fixed context, designates different propositions with respect to different circumstances (possible worlds).

The comparison with non-rigid definite descriptions does not help the suggestion that ~K*p* is non-rigid in the relevant sense. Intuitively, a sentence like 'The number of the planets is less than fifty', as uttered in this context with the definite description understood non-rigidly, designates the same proposition with respect to all circumstances. The variation in the designation of the definite description does not constitute variation in the proposition expressed by a sentence containing it. Indeed, the former variation is best explained by the assumption that the description contributes the same property with respect to all circumstances, coupled with the assumption that different items have that property in different circumstances. If we follow the Russellian tradition and analyse the definite description as a quantifier, then what is rigid is a quantifier—precisely the category of constituent that is supposed to be non-rigid on Kvanvig's proposal. On the Russellian view, the rule of universal instantiation does not permit the substitution of definite descriptions for individual variables (as in the problematic argument about the number of the planets) for the simple reason that definite descriptions are not singular terms.

We do not expect variation in the extension of 'dog', as uttered in a fixed context, with respect to different circumstances of evaluation to constitute variation in the proposition expressed by the sentence 'Fido is a dog'; why should it constitute variation in the proposition expressed by the sentence 'Some dogs bark'? Variation in the extensions of constituent terms of a proposition without variation in the proposition itself is just what is needed for the proposition to have its truth-value contingently.

We have no grounds to suppose that a sentence, as uttered in a fixed context, can designate different propositions with respect to different circumstances of evaluation. The idea that it can seems to confuse expression with evaluation. What a sentence expresses is conceptually prior to the procedure of evaluation, and not relative to a circumstance of evaluation at all.

Kvanvig's proposal assumes that it is contingent what beings or times there are, but threatens our ability to express that very proposition. For suppose that the actual *F*s are exactly $a_1, \ldots, a_n$. If the propositional constituents actually expressed by the *F*-restricted quantifiers $\exists_F$ and $\forall_F$ were tied to $a_1, \ldots, a_n$ in the way proposed, then the proposition actually expressed by the sentence

$$\exists_F x_1 \ldots \exists_F x_n((x_1=a_1 \wedge \ldots \wedge x_n=a_n) \wedge \forall_F y(y=a_1 \vee \ldots \vee y=a_n))$$

('$a_1, \ldots, a_n$ are F and every F is one of them') should be a necessary truth. Consequently, we should expect the proposition actually expressed by its necessitation to be true, and the formula

$$\Diamond{\sim}\exists_F x_1 \ldots \exists_F x_n((x_1=a_1 \wedge \ldots \wedge x_n=a_n) \wedge \forall_F y(y=a_1 \vee \ldots \vee y=a_n))$$

to express a false proposition. But if it is contingent what Fs there are, then the latter sentence should have a true reading. Another kind of quantification appears to be needed to provide that reading.

Kvanvig allows that we could introduce unrestricted rigid quantification over all possible beings and times, but supposes that this interpretation of K would trivialize Fitch's argument by making 'strong' verificationism self-evidently as weak as weak verificationism. That is, he supposes that SVER, read as 'Every proposition, if true, is known at some possible time by some possible being', says no more than WVER, read as 'Every proposition, if true, could be known at some time by some being'. But that is a mistake. Although 'time' and 'being' occur within the scope of 'possible' in the reading of SVER, 'known' does not.

For SVER to be true in the actual world, every truth in the actual world must be known *in the actual world* at some possible time by some possible being. Since actual knowing happens at an actual time by an actual subject, the possibilist liberalization of the quantifiers makes no substantive difference to what the actual truth of SVER requires. Thus SVER is still absurd, and Fitch's argument succeeds as a reductio ad absurdum of WVER. Kvanvig's mistake is like that of confusing 'I am eating something that could have been a cake' with 'I could have been eating a cake'; the first but not the second entails that I am eating. Similarly, even read with possibilist quantifiers, SVER but not WVER entails that every truth is known (not necessarily now).[10]

The use of universal instantiation into sentence position is harmless in Fitch's argument.

12.4 UNANSWERABLE QUESTIONS

Fitch's argument shows that if there are unknown truths then there are unknowable truths. As Joseph Melia (1991) points out, it does not show that if there are unanswered questions then there are unanswerable questions. More precisely, it does not show that if for some proposition *p* it is unknown whether *p* is true then for some proposition *p* it is unknowable whether *p* is true. In particular, if *p* is an unknown truth then it is unknowable *that p* is an unknown truth, but it does not follow that it is unknowable *whether p* is an unknown truth. For that it is an unknowable truth that *p* is an unknown truth does not imply the metaphysical impossibility of a situation in which *p* is false and even known to be false, and thereby known not to be an unknown truth. Equally, that it is an unknowable truth that *p* is an unknown truth does not imply the metaphysical impossibility of a situation in which *p* is known to be true, and even known to be known to be true, and thereby known not to be an unknown truth. In situations of both kinds, it is known whether *p* is an unknown truth. Indeed, Fitch's argument does not show the impossibility of omniscience: a situation *s* such that, for every proposition *p*, it is known in *s* whether *p* is true (in *s* as opposed to actuality). The world might take an especially simple form in *s*, rendering it easier to know; naturally, the cognitive capacities of beings in *s* would also have to be far more extensive than in actuality. The possibility of

[10] See Williamson 1998a, 2000a for more on the possibilist interpretation of the quantifiers.

omniscience would entail that, for every proposition *p*, it can be known whether *p* is true in this weak sense:

WDEC $\forall p\Diamond(\mathrm{K}p \vee \mathrm{K}{\sim}p)$

If WDEC is false, the reasons are different from those explored in this chapter.

Although Fitch's argument does not refute WDEC, it does not follow that we have reason to believe WDEC. In particular, the anti-realist arguments advanced by Michael Dummett, Crispin Wright, and others for the weak verificationist thesis WVER cannot be reinterpreted as arguments merely for the weak decidability thesis WDEC. For the point of those arguments is to identify a difficulty in the supposition that speakers' use of a language sometimes associates a sentence with a truth-condition that can obtain even when they have no disposition to recognize that it obtains. This supposed difficulty is in no way met by the concession that, in some circumstances in which the truth-condition does not obtain, speakers recognize that it does not obtain. For that does not explain why the sentence expresses a truth-condition which does obtain unrecognizably in other circumstances. The anti-realist arguments in question support something like WVER if they support anything at all. If they are not sound arguments for WVER, they are not sound arguments for WDEC either. This is a recurrent problem for responses on behalf of anti-realism to Fitch's argument which modify WVER: they fail to show that anti-realist arguments for WVER are better reinterpreted as arguments for the modified thesis.

12.5 TRANS-WORLD KNOWABILITY

We have examined objections to the soundness of Fitch's argument, and found them defective. Henceforth, we will assume that Fitch's argument is sound. There are unknowable truths in the sense of the formula $\exists p(p \wedge {\sim}\Diamond\mathrm{K}p)$. In this final section we examine attempts to formulate a different sense in which all truths *are* knowable, at least for all Fitch's argument shows, and thereby to mitigate its effect. The upshot will be that the alternative sense is a rather trivial one. But it is not obviously trivial.

The picture naturally associated with the claim that all truths are knowable is that if a truth about some subject matter is unknown, then that epistemic position could be different without any difference in the subject matter itself. Fitch's argument reminds us that we cannot always cleanly separate the unknowing from the unknown in that way, because the epistemic position may be part of the subject matter. Knowing can make a difference to the unknown. Nevertheless, one might object, a difference in the epistemic position makes no difference to how the subject matter was *in the original world in which it was unknown*, as opposed to the world in which the epistemic position is different. In particular, if p is true in this actual world, why should it not be known in some non-actual world w that p is true in this actual world, rather than in w? If we read the operator A as 'in this actual world' or 'actually', we can formulate that idea thus:

WAVER $\forall p(\mathrm{A}p \supset \Diamond \mathrm{KA}p)$

WAVER is in effect the restriction of WVER to instances with the initial operator A. In compensation for the loss of WVER, WAVER claims to provide for each actual truth p a corresponding knowable actual truth Ap. Moreover, one can actually know a priori that p is equivalent to Ap; sentences of the form 'P if and only if in this actual world P' (with their present meaning) are guaranteed not to express untrue propositions. Of course, in the hypothetical world in which Ap is known, its actual equivalence to p may not be known (a priori or a posteriori), otherwise p could be known too in that world, WAVER would yield WVER and Fitch's argument would apply. The phrase 'in this actual world' as uttered in a counterfactual world with its current meaning refers to that counterfactual world, not to this actual one. Only as uttered in this actual world does the sentence 'In this actual world P' expresses the proposition that in this actual world P. But when we actually say 'Someone could have known that in this actual world P', we are using 'in this actual' in this actual world to refer to it, not to a counterfactual world.

Dorothy Edgington (1985) and, less clearly, George Schlesinger (1985: 103–6) proposed just such a modification of weak verificationism in response to Fitch's argument. Edgington generalizes WAVER from the actual world to all possible worlds within a framework of two-dimensional modal logic by prefixing an operator read 'Fixedly', which functions like a necessity operator except that in effect it universally quantifies over worlds of utterance rather than worlds of evaluation. More recently, Wlodek Rabinowicz and Krister Segerberg (1994) have worked out the technical details involved in interpreting WAVER when the knowledge operator K also has its semantics given in terms of possible worlds.

If we try to apply Fitch's argument to WAVER by making the critical substitution of $p \wedge {\sim}\mathrm{K}p$ for p, FACT and DIST get us as far as $\mathrm{A}(p \wedge {\sim}\mathrm{K}p) \supset \Diamond(\mathrm{KA}p \wedge \mathrm{A}{\sim}\mathrm{K}p)$, but there is no obvious absurdity in the consequent, for in some other possible world it may be known that p is true in this actual world, and still be true that in this actual world p is not known. Edgington compares 'actually' to indexicals such as 'I' and 'now'. Other people may know that p is true and I do not know p, although of course they must express their knowledge by saying something like 'p is true and he does not know p', not by saying 'p is true and I do not know p'. Similarly, at future times it may be known that p is true and not known now, although of course one will then have to express that knowledge by saying something like 'p is true and p was not known then', not by saying 'p is true and p is not known now'. Fitch's argument draws no disturbing consequences from WAVER.

WAVER can be strengthened to a biconditional, for its converse is uncontentious. By FACT, $\Diamond KAp$ yields $\Diamond Ap$; the actual truth of $\Diamond Ap$ requires the truth of Ap at some possible world; but since the semantic rule for the operator A evaluates Ap as true with respect to any world if and only if p is evaluated as true with respect to the actual world, that requires the actual truth of p and of Ap. Thus WAVER yields:

WAVER$^+$ $\quad \forall p(Ap \equiv \Diamond KAp)$

WAVER$^+$ may encourage those verificationists who identify truth with knowability, and who therefore require knowability to be sufficient as well as necessary for truth. WVER, by contrast, resists strengthening to a biconditional. For example, suppose that I toss a coin, it comes up heads, and I know that it did. Nevertheless, it could have come up tails, and if it had, I would have known that it had. Thus it could have been known that the coin came up tails; in that sense, it is knowable that it came up tails—even though it did not come up tails.[11] This is a further illustration of the failure in WVER to keep fixed how things are with the subject matter while varying the epistemic position. A verificationist might therefore suspect that the complex operator $\Diamond KA$ comes closer to the intended sense of 'it is knowable that' than does $\Diamond K$ alone.

A curious feature of WAVER and WAVER$^+$ is that they concern knowledge only of necessary truths. Just as the semantics of A validates $\Diamond Ap \supset Ap$, so it validates $Ap \supset \Box Ap$. How things are in this actual world is not contingent; what is contingent is whether this actual world

[11] With respect to a similar problem in defining a priori knowledge, von Wright 1957: 183–4 and, following him, Geach 1972: 181 suggest replacing 'it is possible to know p' by 'there is a way of knowing p'. A different response would be to treat $\Diamond$ as expressing a form of possibility for which a world x is possible with respect to a world w only if the subject matter of the knowing in question does not differ between w and x, but Fitch's argument shows how difficult it is to make sense of that idea. Williamson 1990b: 300–1 has further discussion.

obtains. Fitch's argument reveals a limit to possible knowledge of contingent truths; we may wonder how far its effect is mitigated by possible knowledge of necessary truths.

A deeper point arises. As already noted, counterfactual knowledge of the actual truth of *p* cannot be counterfactually expressed in words such as '*p* is actually true', for counterfactual uses of 'actually' do not refer to the actual world, just as your use of 'I' does not refer to me and past or future uses of 'now' do not refer to now. But then how *could* knowledge of the actual truth of *p* be counterfactually expressed? If this actual world had not obtained, how could anyone have referred to it? If counterfactual knowers could not refer to this actual world, they could not think about it; if they could not think about it, how could they know that *p* is true in it?[12]

In the case of the first person, if you know that I am sitting, then although you cannot express your knowledge by saying 'I am sitting', you can express it by saying 'You are sitting' to me, or by saying 'He is sitting' or 'Williamson is sitting' to someone else. Perhaps you associate 'you', 'he', and 'Williamson' with modes of presentation of me different from the mode of presentation of me that I associate with 'I', but even without the latter mode of presentation of me you can know that I am sitting. Your reference to me typically depends on a causal connection between you and me, mediated by perception and perhaps testimony. Counterfactual knowers cannot refer like that to this actual world, for they are not causally connected to it. One cannot perceive a possible world other than one's own, or receive testimony from it. Similarly, in the case of tense, if on future days we know that it rained today (31 October 1999), although we cannot express our knowledge by saying 'It rained today', we can express it by saying 'It rained yesterday' tomorrow, or by saying 'It rained then' with a memory of today, or simply by saying 'It rained on 31 October 1999'. Perhaps we shall associate 'yesterday' uttered tomorrow, the memory demonstrative 'then', and '31 October 1999' with modes of presentation of today different from the mode of presentation of it that today we associate with today, but even without the latter mode of presentation we can still know on future days that it rained today. Those ways of referring at one time to another are not analogous to ways of referring in one possible world to another possible world. Future reference to today on the basis of a memory demonstrative depends on remembering today, and therefore on a causal connection with today's events; there is no such connection from one possible world to another. Reference to today with 'yesterday'

[12] See also Soames 1998: 13–17 on counterfactual propositional attitudes to actuality.

tomorrow or '31 October 1999' depends on the temporal order, an ordering of times that is independent of what happens at those times. We have no such ordering of worlds. How is counterfactual reference to this actual world to be achieved?

The obvious means of reference to a possible world w in a world other than w is descriptive: one specifies w by specifying what is true at w. Let c be a long conjunction which can be expressed in a counterfactual world x, and is true at this actual world and at no other world. Then perhaps knowers in x could grasp and know Ap by grasping and knowing $\Box(c \supset p)$. Would that solve the problem for WAVER and WAVER$^+$?

The descriptive solution is in one way too cheap, in another too expensive. It is too cheap, for if p is actually true, then it can be included as a conjunct of the long conjunction c describing the actual world. Thus, in counterfactually knowing Ap by having knowledge that one would express as $\Box(c \supset p)$, one has knowledge no more substantive than knowledge of the trivial truth $\Box((p \wedge c) \supset p)$. If that kind of knowledge satisfies WAVER and WAVER$^+$, then they do nothing serious to mitigate the effect of the limits to knowledge which Fitch's argument reveals. In a different respect, the descriptive solution is too expensive, for it requires the counterfactual knower to specify the actual world down to the finest details in the conjunction c, a scarcely imaginable achievement. This way in which the descriptive solution is too expensive hardly compensates for the way in which it is too cheap. Moreover, it is not plausible that counterfactual knowledge of $\Box(c \supset p)$ constitutes knowledge of Ap, for actual knowledge of $\Box(c \supset p)$ hardly constitutes knowledge of Ap, since it does not put one in a position to know p if one is not in a position to know c, and it is unclear why counterfactual knowledge of $\Box(c \supset p)$ should come any closer than actual knowledge of it to knowledge of Ap.

Edgington suggests a different way of specifying counterfactual possibilities: by using counterfactual conditionals. Suppose, for example, that it actually rained last night and no one ever knows that it did. Consider a world w which would have obtained if someone had known that it rained last night. We may also assume that in w some such person knows on reflection that it would still have rained last night even if no one had known that it did, and therefore that if no one had known that it rained last night it would have been an unknown truth that it rained last night ($\mathrm{K}({\sim}\mathrm{K}p \;\Box\!\!\rightarrow (p \wedge {\sim}\mathrm{K}p))$). But the counterfactual supposition that no one knows that it rained last night seems to take us back from w to the actual world; does that not permit us to count the knowledge in w that if no one had known that it rained last night it would have been an unknown truth that it rained last night as knowledge in w that it is actually an unknown truth that it rained last night? More generally, the proposal would be that if, in a world w, another world x would have obtained if the proposition q had been true, then knowledge in w that if q had been true the proposition r would have been true constitutes knowledge in w that r is true in x. Knowers in w need not describe x in detail. They describe one difference (q) from their world and let everything else be as close as possible to their world in the manner of counterfactual suppositions; they need not even know what their world is like. In particular, x can be the actual world, q the proposition $\sim Kp$, and r the proposition $p \wedge \sim Kp$.

Unfortunately for the proposal, the descriptive element in the antecedent of the counterfactual suffices to regenerate the trivialization argument. For suppose that, in the world w, the world x would have obtained if q had been true, and that r is true in x. Then, in w, x would have obtained if the conjunction $q \wedge r$ had been true; in the terms of a simple possible worlds semantics for the counterfactual conditional, if r is true in x and x is the closest world to w in which q is true then x is the closest world to w in which $q \wedge r$ is true. The proposal therefore implies that knowledge in w of the counterfactual $(q \wedge r) \,\Box\!\!\rightarrow r$ constitutes knowledge in w that r is true in x. But since r is a truth-functional consequence of $q \wedge r$, the counterfactual $(q \wedge r) \,\Box\!\!\rightarrow r$ is a trivial necessary truth. Knowledge of it in w is knowledge of nothing interesting.

It would not help to stipulate that the relevant counterfactual not be a logical truth. For let s state something contingent but utterly outlandish, logically quite independent of both q and r, such that it is obvious in w that there are much closer worlds to w in which $q \wedge r$ is true than any in which s is true. Then, as before, x is the closest world to w in which the disjunction $(q \wedge r) \vee s$ is true. Thus the counterfactual $((q \wedge r) \vee s) \,\Box\!\!\rightarrow r$ is true but not logically true in w, so knowledge of it constitutes knowledge in w that r is true in x even by the modified proposal. But knowledge of the counterfactual $((q \wedge r) \vee s) \,\Box\!\!\rightarrow r$ is still trivial by contrast with knowledge of Ar, because its basis is just that s is a far more outlandish supposition than $q \wedge r$. Rabinowicz and Segerberg admit that they are not sure that the trivialization problem can be overcome (1994: 113–14).

The specification of worlds by counterfactual conditionals also faces a problem of specificity, for why should we suppose that there is a unique closest world to w in which q is true? For example, why should the counterfactual supposition in w that no one ever knows that it rained last night single out the actual world uniquely? David Lewis (1973) rejects the uniqueness assumption in his semantics for counterfactual conditionals. We could make the looser requirement that, in w, x is one of the worlds which might have obtained if q had been true; but if the counterfactual supposition made by the knowers in w does not single out x uniquely, then it is not obviously legitimate to characterize them as knowing something specifically about x—for example, about the actual world.

The problem of specificity is implicit in the formulations WAVER and WAVER$^+$ themselves. For $\sim Ap \equiv A\sim p$ is a theorem of standard logics of the operator A; either p is actually true or $\sim p$ is. Consequently, any counterfactual knower who can somehow express the propositional constituent which we express with A is in a position to specify this actual world; it is the world with respect to which the proposition $\forall p(p \equiv Ap)$ is true. Thus reference to a complete world is built into the technical conception of actuality to which Edgington appeals. Moreover, we can use 'actual' to make comparatively unproblematic reference because our ostension takes place in a unique complete world, *this* world. Incomplete worlds differ from complete ones in not being mutually exclusive. If we try to use 'actual' to refer to an incomplete world, our ostension with '*this* world' takes place in many worlds of varying degrees of incompleteness simultaneously; which of them is being ostended?

These considerations suggest that the 'actually' operator is not best suited to Edgington's purposes. She herself says 'Knowledge of counterfactual situations is never of one specific possible world' (1985: 564). Her preferred formulations are in terms of unspecific possible situations. We can adapt WAVER in this way, generalizing it beyond the actual world:

$$\text{WSVER} \quad \forall p \forall s(\text{In}(s,p) \supset \exists s^*\text{In}(s^*,\text{KIn}(s,p)))$$

This says in effect that for every possible situation s, if p is true in s then in some possible situation s^* it is known that p is true in s. Sten Lindström (1997) has worked out the technical details of such a situation-theoretic approach.

A radical version of situation theory *forbids* situations to involve the whole universe. If so, then it cannot be true *in* any situation that p is an unknown truth, for 'unknown' abbreviates the unrestricted generalization 'not known by anyone at any time', which on its intended understanding involves the whole universe. That version of the approach scarcely engages with Fitch's argument; it restricts verificationism to statements about limited portions of the universe. If p is an unknown truth in a limited situation s, where the quantifiers implicit in 'unknown' are restricted to s, then an observer outside s may be able to perceive that *p* is an unknown truth in *s*, as Lindström points out (1997: 197). It is unclear how semantic arguments for verificationism could be reinterpreted to yield only that limited version of it, unless—very implausibly—they denied the intelligibility of unrestricted quantification over all subjects and all times.

A moderate version of situation theory permits but does not require situations to involve the whole universe. Thus it can be true in some situation *s* that *p* is an unknown truth, in a sense from which it follows that if *s* obtains then *p* is an unknown truth *simpliciter*. Then, according to WSVER, it is known in some possible situation *s** that *p* is an unknown truth in *s*. It is not immediate that *s* and *s** are not compossible. If they both obtained, might someone in *s** know that *p* is an unrestrictedly unknown truth in *s* without knowing that *p* is an unrestrictedly unknown truth, by not realizing that *s* obtains? Someone who specifies *s* simply by a perceptual or memory demonstrative or by spatio-temporal coordinates without drawing the consequence that *s* obtains is still in a position to know that *s* obtains. We can understand 'unknown' broadly enough to exclude such cases. The knowers in *s** must specify *s* by means which do not require *s* to obtain. But then they must specify *s* by something like description or a counterfactual supposition. Given that *s* is not a complete possible world, the specificity problem is no longer pressing. Perhaps it can be described by a short conjunction; perhaps in *s** it is the most specific situation that would obtain if *p* were not known. However, the trivialization problem is just as serious as before. Knowers in *s** can build *p* as a conjunct into their specification of *s*, and thereby achieve trivial knowledge that *p* is true in *s*; the argument runs just as before. WSVER, even if true, does not seriously mitigate the effect of Fitch's argument. The knowledge that it claims to be possible is achievable in a trivial way if achievable at all.

We may briefly note some further problems for the counterfactual strategy used to defend WSVER; similar criticisms can be made of WAVER and WAVER$^+$.

(i) Edgington's treatment of the proposition that *p* is an unknown truth assumes that *p* itself is knowable in the original sense ($\Diamond$K) in which it is conceded that not all truths are knowable. But *p* may not be knowable in that sense. For example, let *p* be the conjunction of a complete description of all actual neurophysiological events at all times conjoined with the proposition that there are no non-physical thinkers. We may assume that *p* is unknown. If *p* were known, it would be true, and therefore known by a physical knower; but then some neurophysiological events in the brain of that knower would be different from all actual neurophysiological events, so *p* would not be true after all, and so would not be known. Thus *p* is arguably unknowable, because the supposition that *p* is known is self-defeating. But then we could not know that *p* is an unknown truth in an actual situation *s* by knowing *p* in a non-actual situation *s** and reflecting that *p* would still have been true if it had not been known, for *p* cannot be known in any possible situation.

(ii) Now suppose that *p* is a knowable but unknown truth in *s*. The idea is that if *s** is the closest situation to *s* in which *p* is known, then *s* is the closest situation to *s** in which *p* is unknown, so subjects in *s** can specify *s* as the situation which would obtain if *p* were unknown. But the relation between *s* and *s** need not be so symmetrical. For example, let *p* be the proposition that there is a pebble at spatio-temporal location *xyzt*, and *s* be a situation in which *p* is true but unknown because the conditions for intelligent life emerge only long after *t* (the time of *xyzt*). Let *s** be a situation as close as possible to *s* in which *p* is known. Cosmic history follows vastly divergent paths in *s* and *s**. In the closest possible situations to *s** in which *p* is unknown, it is unknown simply because no one chances to travel near *xyzt*; such situations are far closer to *s** than to *s* in cosmic history. Thus although someone in *s** may know that if *p* had been unknown it would have been an unknown truth, that cannot be represented as knowledge in *s** that *p* is an unknown truth in *s*, for, in *s**, if *p* had been unknown *s* would not have obtained. Knowers in *s** may be unable to single out *s* by any counterfactual supposition. Although the knowledge that WSVER attributes is achievable in trivial ways if achievable at all, it may not be achievable at all.

(iii) Edgington's counterfactual strategy makes another assumption: that if a proposition *p* is an unknown truth in a situation *s*, and *s** is the closest situation to *s* in which *p* is known, then, in *s**, if *p* had not been known it would still have been true. What happens if that assumption breaks down? Edgington argues:

> If there are truths which fail to satisfy the principle
>
> (7) *p* would still have been true, had no one known that *p*,
>
> that I am in pain, for example, then they satisfy the principle,
>
> (8) If *p*, then someone knows that *p*
>
> and, *a fortiori*, they satisfy [WAVER]. (1985: 567; numbering added)

There is a lacuna here, for the counterfactual strategy requires (7) to be known, not just true, in *s**; the observer-independence of what we observe is not always easy to establish. Because I wonder whether asking myself whether I am embarrassed causes me to be embarrassed, I may not know that I would still have been embarrassed, had no one known that I was embarrassed (I may know that no one else will ever know that I was embarrassed). Thus even if the falsity of (7) implies the truth of (8), that does not help when (7) is true but unknown, and perhaps even unknowable (in the sense in which Fitch's argument shows that there are unknowable truths).

(iv) But does the falsity of (7) imply the truth of (8) in the relevant situation? The difficulty for the counterfactual strategy arises when (7) is false in *s**. If the falsity of (7) implies the truth of (8), then what follows is that (8) is true in *s**. That is no surprise, for the consequent of (8) is true in *s**. But since it is consistent with the truth of (8) in *s** that *p* is an unknown truth in *s*, no way has yet been provided of knowing either in *s** or in *s* that *p* is an unknown truth in *s*. Thus WSVER is still under threat. An argument to the truth of (8) in *s* from the falsity of (7) in *s* would be unlikely to help, for, by hypothesis, both the antecedent and the consequent of (7) are true in *s*; (7) is probably true in *s*. What is really needed is an argument to the truth of (8) in *s* from the falsity of (7) in *s**. But there is no such connection. Suppose that our best physical theory tells us that one can know what state a ?-particle is in only by interacting with it in ways which unpredictably change its state; one knows what state it is in after the change. Suppose also that it is an unknown truth in *s* that a ?-particle *z* is in a state *k* (that proposition is *p*). Then (7) is not true in the corresponding situation *s** in which it is known that *z* is in state *k*, for if it had not been known that *z* was in *k*, the interactions would not have occurred and *z* might well not have been in *k*. Nevertheless, (8) is not true in *s*. Of course, a verificationist might deny that there can be *unknown* truths in cases of this kind: but such a claim has no plausibility for those not antecedently committed to verificationism. At any rate, we have no good reason to think that the falsity of (7) in *s** entails the truth of (8) in *s*. That is another respect in which the counterfactual strategy is insufficiently general.

(v) The counterfactual strategy is inadequate as a defence of WAVER and WSVER, and the knowledge which they ascribe is anyway trivial in the sense explained above. There is also an underlying problem about the motivation for such modified verificationist principles. A verificationist principle (WVER) was originally motivated by arguments about the nature of meaning. In response to Fitch's argument, the principle was modified. But it was not checked that the meaning-theoretic argument for WVER could plausibly be reconstrued as an argument for WAVER or WSVER. Such a reconstrual looks quite dubious. For the meaning-theoretic argument for WVER proceeds at the level of sentences; roughly, it attempts to demonstrate that the truth-condition of a sentence somehow depends on its assertibility-condition in such a way that if the assertibility-condition is impossible (incapable of obtaining), then so too is the truth-condition. Sentences of the form 'A and it is not known that A' constitute counterexamples to that claim. Given a sentence S with a possible truth-condition and an impossible assertibility-condition, the defence of WAVER or WSVER, if successful, would produce a complex sentence Φ(S) of which S is a constituent, and show that Φ(S) has a possible assertibility-condition. For example, if S is 'It rained last night without its being known that it rained last night', Φ(S) might be 'If it had not been known that it rained last night, it would have been the case that it rained last night without its being known that it rained last night'. Presumably, the idea is that the possibility of the assertibility-condition of Φ(S) somehow explains the possibility of the truth-condition of S in a manner consistent with the spirit of a verificationist theory of meaning. Since the truth-condition of S is an aspect of the meaning of S, even on a verificationist account, the meaning of S is in effect being explained in terms of the meaning of Φ(S). But verificationist theories of meaning of the kind used in the arguments for WVER, WAVER, and WSVER are compositional; they give the meaning of complex expressions in terms of the meanings of their constituents. Since S is a constituent of Φ(S), the meaning of Φ(S) is explained in terms of the meaning of S. This looks like an explanatory circle. Perhaps the defender of WAVER or WSVER has some way of rendering it harmless, but we should not rush to assume that the defence of those principles can be reconciled with the meaning-theoretic ideas which were supposed to motivate the original weak verificationism. Sometimes we should learn from counterexamples that a philosophical idea was wrong in spirit, not just in letter.[13]

Point (i) above indicated the presence of unknowable truths similar in spirit but formulated without use of specifically epistemological concepts. The domain of unknowability is probably far wider than that. Once we acknowledge that the domain is non-empty, we can explore more effectively its extent. In order to be able to set a limit to knowledge, we do not have to find both sides of the limit knowable. Although, trivially, we cannot know that which we cannot know, we

[13] See Percival 1991 and Wright 1993: 428–32 for further discussion of Edgington's proposal.

can know that we cannot know something. In this chapter we saw a route to knowing of various pairs of propositions that since both are unknowable and one or other of them is true, one or other of them is an unknowable truth. What we have not seen is a route to knowing that when the pair consists of a proposition and its negation (see section 12.4). Yet we may plausibly conjecture that, in some sense of 'impossible', we can know of some propositions both that they are true or false and that it is impossible to know them to be true and impossible to know them to be false. We are only beginning to understand the deeper limits of our knowledge.

APPENDIX 1
Correlation Coefficients

Given real-valued random variables X and Y whose arguments are conditions ('events') in a suitable probability space, their *correlation coefficient* $\rho[X,Y]$ is defined as $\mathrm{Cov}[X,Y]/(\sigma[X]\sigma[Y])$; their covariance $\mathrm{Cov}[X,Y]$ is in turn defined as $E[XY]-E[X]E[Y]$, where $E[X]$ is the expectation of X; the standard deviation $\sigma[X]$ is defined as $\sqrt{}E[(X-E[X])^2]$ (e.g. Parzen 1960: 362). Covariance is not itself an adequate measure of correlation; for example, any random variable is perfectly correlated with itself, but $\mathrm{Cov}[X,X] = \sigma[X]^2$, which varies. One must calibrate $\mathrm{Cov}[X,Y]$ by dividing by $\sigma[X]\sigma[Y]$.

All these notions can be relativized to a set of background conditions by conditionalizing the underlying probabilities on those conditions.

We must now define correlation coefficients for the conditions on which the probabilities are defined. Denote the probability of a condition C as $P[C]$. The indicator random variable $X(C)$ is 1 if C obtains and 0 otherwise; thus $E[X(C)] = P[C]$. For any conditions C and D we define their correlation coefficient $\rho[C,D]$ (with harmless ambiguity in ρ) as $\rho[X(C),X(D)]$. We will calculate an expression for $\rho[C,D]$ in terms of probabilities, and then derive some elementary facts about it. The conditional probability $P[C|D]$ is $P[C \wedge D]/P[D]$ $(0 < P[D])$.

We assume that $0<P[C]<1$, and similarly for D and E, otherwise the correlation coefficients for these states are not well-defined.

Proposition 1. $\rho[C,D] = (P[C \wedge D]-P[C]P[D])/\sqrt{(P[C](1-P[C])P[D](1-P[D]))}$.

Proof.
$$\begin{aligned}
\mathrm{Cov}[X(C),X(D)] &= E[X(C)X(D)]-E[X(C)]E[X(D)]\\
&= E[X(C \wedge D)]-E[X(C)]E[X(D)]\\
&= P[C \wedge D]-P[C]P[D].\\
\text{Moreover, } \sigma[X(C)] &= \sqrt{}E[(X(C)-E[X(C)])^2]\\
&= \sqrt{(E[X(C)^2]-(E[X(C)])^2)} \quad \text{by a standard calculation}\\
&= \sqrt{(E[X(C)]-(E[X(C)])^2)} \quad \text{because } X(C)\in\{0,1\}\\
&= \sqrt{(P[C](1-P[C]))}. \ \blacksquare
\end{aligned}$$

Proposition 2. (a) If $P[C|D] > P[C]$ then $\rho[C,D] > 0$.
(b) If $P[C|D] = P[C]$ then $\rho[C,D] = 0$.
(c) If $P[C|D] < P[C]$ then $\rho[C,D] < 0$.

Proof. By Proposition 1, $\rho[C,D]$ has the same sign as $P[C \wedge D]-P[C]P[D] = (P[C|D]-P[C])P[D]$. $\blacksquare$

Proposition 3. (a) If $\rho[C,D]\geq 0$ and $\rho[C,E]\geq 0$ then $\rho[C,D]\leq\rho[C,E]$ iff $(P[C|D]-P[C])(P[C]-P[C|{\sim}D]) \leq (P[C|E]-P[C])(P[C]-P[C|{\sim}E])$

(b) If $\rho[C,D]\leq 0$ and $\rho[C,E]\leq 0$ then $\rho[C,D]\leq\rho[C,E]$ iff $(P[C|D]-P[C])(P[C]-P[C|{\sim}D]) \geq (P[C|E]-P[C])(P[C]-P[C|{\sim}E])$

Proof. (a)
$$\begin{aligned}(P[C|D]-P[C]P[D])^2 &= (P[C|D]P[D]-P[C]P[D])(P[C]-P[C\wedge{\sim}D]-P[C]P[D])\\ &= (P[C|D]-P[C])P[D](P[C](1-P[D])-P[C|{\sim}D]P[{\sim}D])\\ &= (P[C|D]-P[C])(P[C]-P[C|{\sim}D])P[D](1-P[D]).\end{aligned}$$
Hence by Proposition 1,
$$\rho[C,D]^2P[C](1-P[C]) = (P[C|D]-P[C])(P[C]-P[C|{\sim}D]).$$
But if $\rho[C,D]\geq 0$ and $\rho[C,E]\geq 0$ then
$$\begin{aligned}\rho[C,D] \leq \rho[C,E] &\Leftrightarrow \rho[C,D]^2P[C](1-P[C])\leq\rho[C,E]^2P[C](1-P[C])\\ &\Leftrightarrow (P[C|D]-P[C])(P[C]-P[C|{\sim}D])\\ &\qquad \leq (P[C|E]-P[C])(P[C]-P[C|{\sim}E]).\end{aligned}$$

(b) is similar. ■

Proposition 4. If $P[C|D]\leq P[C|E]$ and $P[C|{\sim}D]\geq P[C|{\sim}E]$ then $\rho[C,D]\leq\rho[C,E]$.

Proof. Suppose that $P[C|D]\leq P[C|E]$ and $P[C|{\sim}D]\geq P[C|{\sim}E]$.

Case (i): $P[C]\leq P[C|D]$. Then $P[C]\leq P[C|E]$, so by Proposition 2, $\rho[C,D]\geq 0$ and $\rho[C,E]\geq 0$. Hence by Proposition 3(a), $\rho[C,D]\leq\rho[C,E]$ iff

(*) $(P[C|D]-P[C])(P[C]-P[C|{\sim}D]) \leq (P[C|E]-P[C])(P[C]-P[C|{\sim}E])$.

But $0 \leq P[C|D]-P[C] \leq P[C|E]-P[C]$. Moreover, since $P[C] = P[D]P[C|D] + (1-P[D])P[C|{\sim}D]$, $0 \leq P[C]-P[C|{\sim}D] \leq P[C]-P[C|{\sim}E]$. Hence (*) holds, so $\rho[C,D] \leq \rho[C,E]$.

Case (ii): $P[C|D]<P[C]<P[C|E]$. Then $\rho[C,D]<0<\rho[C,E]$ by Proposition 2.

Case (iii): $P[C|E]\leq P[C]$. Similar to case (i), using Proposition 3(b). ■

Proposition 5. (a) $\rho[C,C] = 1$

(b) $\rho[C,{\sim}C] = -1$

(c) $-1 \leq \rho[C,D] \leq 1$

(d) $\rho[C,D] = 1$ iff $P[C|D] = 1$ and $P[C|{\sim}D] = 0$.

(e) $\rho[C,D] = -1$ iff $P[C|D] = 0$ and $P[C|{\sim}D] = 1$.

Proof. (a) and (b) are simple consequences of Proposition 1.

(c) $P[C|D]\leq P[C|C]$ and $P[C|{\sim}D] \geq P[C|{\sim}C]$, so $\rho[C,D]\leq\rho[C,C] = 1$ by (a) and Proposition 4. Similarly, $\rho[C,D]\geq\rho[C,{\sim}C] = -1$.

(d) Suppose that $\rho[C,D] = 1$. Then $\rho[C,C] \leq \rho[C,D]$ by (a) and (c), and $P[C|D] >P[C]$ by Proposition 2. Hence by Proposition 3

$$(P[C|D]-P[C])(P[C]-P[C|{\sim}D]) \geq (P[C|C]-P[C])(P[C]-P[C|{\sim}C]) = (1-P[C])P[C].$$

But $P[C|D]-P[C]\leq 1-P[C]$ and $P[C]-P[C|{\sim}D]\leq P[C]$, so $P[C|D]-P[C] = 1-P[C]$ and $P[C]-P[C|{\sim}D] = P[C]$, so $P[C|D] = 1$ and $P[C|{\sim}D] = 0$. Conversely, if $P[C|D] = 1$ and $P[C|{\sim}D] = 0$ then $P[D] = P[C \wedge D] = P[C]$ and the result follows by Proposition 1.

(e) is similar to (d). ■

APPENDIX 2
Counting Iterations of Knowledge

We can treat the propositional modal logic KT (=T) as an idealized logic for knowledge by reading the necessity operator as 'One knows that', here written K. As the axioms of KT we use all truth-functional tautologies and all instances of $\mathrm{KA} \supset \mathrm{A}$ ('knowledge implies truth'). As rules we use modus ponens and the rule RK that if $[\mathrm{A}_1 \wedge \ldots \wedge \mathrm{A}_n] \supset \mathrm{B}$ is a theorem, so is $[\mathrm{KA}_1 \wedge \ldots \wedge \mathrm{KA}_n] \supset \mathrm{KB}$ ($n \geq 0$, 'knowledge is closed under logical consequence', with the analogue of the rule of necessitation for $n = 0$). See Chellas (1980) for further details.

Suppose that the tree is in fact k inches tall. We are interested in how many iterations of knowledge one can have of the propositions that it is not j inches tall, where $j<k$, and that if it is n inches tall then one does not know that it is not $n-1$ inches tall, where $j<n\leq k$. Let us read the propositional variable p_i as 'The tree is i inches tall'. For iterations of knowledge we define K^n inductively: $\mathrm{K}^0\mathrm{A} = \mathrm{A}$, $\mathrm{K}^{n+1}\mathrm{A} = \mathrm{K}^n\mathrm{KA}$. For each natural number n from j to k, let $i[n]$ be a natural number. Now consider this formula:

$$(*) \quad \mathrm{K}^{i[j]}\sim p_j \;\wedge\; \bigwedge_{j<n\leq k} \mathrm{K}^{i[n]}(p_n \supset \sim\mathrm{K}\sim p_{n-1}) \wedge p_k$$

According to $*$: one has $i[j]$ iterations of knowledge that the tree is not j inches tall; for each n from j to k, $i[n]$ iterations of knowledge that if it is n inches tall then one does not know that it is not $n-1$ inches tall; the tree is k inches tall. Our question is: for which numbers is $*$ consistent in KT? The answer turns out to be that it is consistent if and only if $i[n]<k-n$ for some n ($j\leq n\leq k$). Thus one can have arbitrarily many iterations of knowledge of all but one of the propositions, provided that one falls below the specified limit for the remaining proposition.

Proposition. For all natural numbers j, k ($j\leq k$) and $i[j], i[j+1], \ldots, i[k]$:
$\vdash_{\mathrm{KT}} \sim *$ if and only if for all n, if $j\leq n\leq k$ then $i[n]\geq k-n$.

Proof. First suppose that for all n, if $j\leq n\leq k$ then $i[n]\geq k-n$. Consider this formula:

$$(**) \quad \mathrm{K}^{k-j}\sim p_j \;\wedge\; \bigwedge_{j<n\leq k} \mathrm{K}^{k-n}(p_n \supset \sim\mathrm{K}\sim p_{n-1}) \;\wedge\; p_k$$

Since $\vdash_{\mathrm{KT}} \mathrm{KA} \supset \mathrm{A}$, $\vdash_{\mathrm{KT}} \mathrm{K}^{i[j]}\sim p_j \supset \mathrm{K}^{k-j}\sim p_j$ and $\vdash_{\mathrm{KT}} \mathrm{K}^{i[n]}(p_n \supset \sim\mathrm{K}\sim p_{n-1}) \supset \mathrm{K}^{k-n}(p_n \supset \sim\mathrm{K}\sim p_{n-1})$ for all n ($j<n\leq k$) by supposition. Hence $\vdash_{\mathrm{KT}} * \supset **$, so it suffices to show that $\vdash_{\mathrm{KT}} \sim**$. Now $\vdash_{\mathrm{KT}} (\mathrm{K}\sim p_{n-1} \wedge (p_n \supset \sim\mathrm{K}\sim p_{n-1})) \supset \sim p_n$ ($j<n\leq k$), so by $k-n$ applications of RK, $\vdash_{\mathrm{KT}} (\mathrm{K}^{k-n+1}\sim p_{n-1} \wedge \mathrm{K}^{k-n}(p_n \supset \sim\mathrm{K}\sim p_{n-1})) \supset \mathrm{K}^{k-n}\sim p_n$, so $\vdash_{\mathrm{KT}} ** \supset (\mathrm{K}^{k-n+1}\sim p_{n-1} \supset \mathrm{K}^{k-n}\sim p_n)$. Putting all these pieces together for n from $j+1$ to k, $\vdash_{\mathrm{KT}} ** \supset (\mathrm{K}^{k-j}p_j \supset \sim p_k)$. But $\vdash_{\mathrm{KT}} ** \supset (\mathrm{K}^{k-j}p_j \wedge p_k)$, so $\vdash_{\mathrm{KT}} \sim**$.

For the converse, suppose that for some h, $j \leq h \leq k$ and $i[h] < k-h$. We use the well-known fact that any KT theorem is true at any world in any standard possible worlds model with a reflexive accessibility relation to show that not $\vdash_{KT} \sim *$, by constructing such a model with a world at which $*$ is true. The worlds are the natural numbers from h to k inclusive; for any worlds m and n, n is accessible from m if and only if $|m-n| \leq 1$. Thus for any formula A and world m, KA is true at m if and only if for every world n, if $|m-n| \leq 1$ then A is true at n. Consequently, for any natural number l, $K^l A$ is true at m if and only if for every world n, if $|m-n| \leq l$ then A is true at n. For $j \leq m \leq k$ and $h \leq n \leq k$, we set p_m true at n if and only if $m = n$. We will show that $*$ is true at k. Certainly p_k is true at k. There are two cases to consider.

(i) $h = j$. Thus for any world n, if $|k-n| \leq i[h]$ then $|k-n| < k-h$ because by hypothesis $i[h] < k-h$, so $h < n$, so $\sim p_h$ is true at n. Consequently, $K^{i[h]} \sim p_h$ is true at k. If $j < n \leq k$, p_{n-1} is true at $n-1$, so $\sim K \sim p_{n-1}$ is true at n; since p_n is true only at n, $p_n \supset \sim K \sim p_{n-1}$ is true at every world; thus $K^{i[n]}(p_n \supset \sim K \sim p_{n-1})$ is true at k.

(ii) $h > j$. If $j \leq n < h$ then p_n is false at every world, so $K^{i[j]} \sim p_j$ and $K^{i[n]}(p_n \supset \sim K \sim p_{n-1})$ are true at every world. $K^{i[h]}(p_h \supset \sim K \sim p_{h-1})$ is true at k because if $|k-n| \leq i[h]$ then $h < n$, so p_h is false at n. If $h < n \leq k$ then $p_n \supset \sim K \sim p_{n-1}$ is true at every world, so $K^{i[n]}(p_n \supset \sim K \sim p_{n-1})$ is true at k. ■

Evidently, this is just a specimen result. We can generalize it by adding as extra conjuncts to $*$ any formulas true at world k in the models specified in the second half of the proof. For example, since $K^i \sim p_n$ is true at k whenever $|k-n| > i$, we can add any such formula. Similarly, one can add $K^i \sim (p_m \wedge p_n)$ for any i, m and n where $m \neq n$. One can also add worlds $k+1$, $k+2$, . . ., subject to the same truth definitions. What is crucial is omitting j as a world when enough iterations of knowledge of $\sim p_j$ are needed.

The accessibility relation just used is symmetric as well as reflexive. The model therefore validates the Brouwersche schema $A \supset K \sim K \sim A$, the addition of which to KT gives the system KTB. KTB is exactly the logic for knowledge determined by the simplest version of the margin for error considerations. The only logical features essential to the binary similarity relation are reflexivity and symmetry, accessibility in the model plays the role of similarity, and KTB is the logic determined by the constraints of reflexivity and symmetry on accessibility. Section 5.3 also describes more sophisticated versions of the margin for error considerations on which the width of the margin varies from point to point, which makes the accessibility relation non-symmetric.

A feature of both KT and KTB is that, for any formula A, $A \supset KA$ is a theorem if and only if either $\sim A$ is a theorem or A is (Williamson 1992a). This is a formal analogue of the hypothesis in Chapter 4 that only trivial conditions are luminous.

APPENDIX 3
A Formal Model of Slight Insensitivity Almost Everywhere

We will use a formal model to study how the slightest systematic inaccuracy can make one almost totally insensitive in the counterfactual sense: for every contingent proposition p, if p were false one might still believe p.

For simplicity, we evaluate counterfactual propositions according to Lewis's semantics, and imagine possible worlds as varying along a single dimension. We treat propositions as sets of worlds. We represent the worlds by the real numbers, and measure the distance between worlds w and x in the natural way, by $|w-x|$. Thus $q \mathrel{\Box\!\!\rightarrow} r$ is true at w if and only if either q is true at no world or q is true at some world x such that for every world y, if $|w-y|\leq|w-x|$ and q is true at y then r is true at y. The same negative results about sensitivity will follow from the weaker assumption that Lewis's condition is necessary for the truth of the counterfactual. For any proposition p, Bp is the proposition that one believes p. One believes p sensitively at w if and only if B$p \wedge (\sim p \mathrel{\Box\!\!\rightarrow} \sim$B$p)$ is true at w.

Let one's process of belief formation embody a very slight bias towards one world, which may as well be 0. More precisely, given a small quantity $\varepsilon>0$, define a mapping f from worlds to worlds:

$$\begin{array}{ll} f(w) = w+\varepsilon & \text{if } w<-\varepsilon \\ f(w) = 0 & \text{if } -\varepsilon\leq w\leq\varepsilon \\ f(w) = w-\varepsilon & \text{if } \varepsilon<w \end{array}$$

Thus f is a shift of at most ε in the direction of 0. To formalize the bias, suppose that for every proposition p and world w, Bp is true at w if and only if p is true at $f(w)$. At 0, one in effect believes one to be in 0 ($f(0) = 0$). Elsewhere, the world in which one believes oneself to be is very slightly closer than the world in which one is to 0. For example, if a is a real number, let $p(a)$ be the proposition true at exactly the worlds w such that $-a<w<a$. Then, for $a>\varepsilon$, B$p(a)$ is true at w if and only if $-a-\varepsilon<w<a+\varepsilon$, so one believes $p(a)$ falsely at w if and only if $-a-\varepsilon<w\leq-a$ or $a\leq w<a+\varepsilon$. Since one believes $p(a)$ at any world at which it is true, the worlds at which one believes $p(a)$ falsely form a tiny proportion of the worlds at which one believes $p(a)$, if ε is small compared to a. The model makes one's beliefs consistent (for p and $\sim p$ are not both true at $f(w)$), complete (for either p is true at $f(w)$ or $\sim p$ is)) and deductively closed (for if q logically follows from $p_1, \ldots, p_{n,}$ and $p_1, \ldots, p_n$ are true at $f(w)$, then so is q) at every world w. These highly idealized properties ensure that any cognitive problems

one has in the model are not the result of logical incapacity. At 0, one believes p if and only if p is true. But one's beliefs are insensitive. For example, $\sim p(a) \mathbin{\Box\!\!\rightarrow} p(a+\varepsilon)$ is true at 0 because if one's world were not within a distance a of 0 it would be within $a+\varepsilon$ of 0; but $\mathrm{B}p(a)$ is true at exactly the same worlds as $p(a+\varepsilon)$, so $\sim p(a) \mathbin{\Box\!\!\rightarrow} \mathrm{B}p(a)$ is true at 0 whenever $\varepsilon>0$. If $p(a)$ were false one would still believe it. Thus one believes $p(a)$ insensitively at 0, in virtue of the tiny belt of worlds at which one believes it falsely between the broad core of worlds at which one believes it truly and the infinitely broad remainder of worlds in which it is false and one disbelieves it.

The point generalizes: one believes *no* contingent proposition sensitively at any world within a distance $\varepsilon/2$ of 0. Proof: Suppose that $\mathrm{B}p \wedge (\sim p \mathbin{\Box\!\!\rightarrow} \sim \mathrm{B}p)$ is true at w, where $|w|\leq\varepsilon/2$ and p is contingent. Without loss of generality we may assume that $0\leq w\leq\varepsilon/2$. Since p is contingent, the truth of the counterfactual at w requires a world x at which $\sim p$ is true and for every world y, if $|w-y|\leq|w-x|$ then $\sim p \supset \sim \mathrm{B}p$ is true at y. Let $f^0(x) = x$ and $f^{n+1}(x) = f(f^n(x))$. We show by induction on n that p is false at $f^n(x)$ and $|w-f^n(x)|\leq|w-x|$ for all n. Basis: Trivial. Induction step: By the induction hypothesis, $\sim p \supset \sim \mathrm{B}p$ is true and p false at $f^n(x)$, so $\mathrm{B}p$ is false at $f^n(x)$, so p is false at $f^{n+1}(x)$. We must show that $|w-f^{n+1}(x)| \leq |w-x|$. If $f^n(x)\leq 0$ then $f^n(x)\leq f^{n+1}(x)\leq 0\leq w$, so $|w-f^{n+1}(x)|\leq|w-f^n(x)| \leq |w-x|$ by the induction hypothesis. If $3\varepsilon/2\leq f^n(x)$ then $w\leq\varepsilon/2\leq f^{n+1}(x)\leq f^n(x)$, so again $|w-f^{n+1}(x)|\leq|w-f^n(x)|\leq|w-x|$. If $0<f^n(x)<3\varepsilon/2$ then $0\leq f^{n+1}(x)<\varepsilon/2$, so $|w-f^{n+1}(x)|\leq\varepsilon/2$; since $\mathrm{B}p$ is true at w, p is true at $f(w)=0$; but $\mathrm{B}p$ is false at x, so p is false at $f(x)$, so $f(x)\neq 0$, so $\varepsilon<|x|$, so $\varepsilon/2\leq|w-x|$, so $|w-f^{n+1}(x)|\leq|w-x|$. This completes the induction. By definition of f, $f^n(x)=0$ for some n. Hence p is false at 0, which yields a contradiction. Thus no contingent proposition is believed sensitively at w if $|w|\leq\varepsilon/2$. ■

Contingent propositions are believed sensitively at worlds more distant from 0. If $\varepsilon/2<w<\varepsilon$, a proposition false at $(2w-\varepsilon)/3$ and $2(w+\varepsilon)/3$ and true elsewhere is believed sensitively at w. If $\varepsilon\leq w$, a proposition false at $w-2\varepsilon/3$ and $w+\varepsilon/3$ and true elsewhere is believed sensitively at w.

One's only sensitive beliefs at worlds near 0 are beliefs in necessary propositions, which are vacuously sensitive. Given that necessary propositions entail only necessary propositions, (4) in section 7.5 implies that at these worlds one has *no* knowledge of contingencies if there is any such bias at all, no matter how small. As soon as $\varepsilon = 0$, all beliefs are sensitive and may count as knowledge. Is a notion of sensitivity on which the slightest bias can produce total insensitivity an adequate basis for an account of knowledge attribution? Intuitively, we might expect a slight bias to rule out knowing p close to the boundary of worlds at which p is true; on a sensitivity-based account, it can do so arbitrarily far from the boundary.

We can modify the model to give it more desirable features. For example, as it stands it assigns one some Moore-paradoxical beliefs at some worlds (although not at 0). Since $p(\varepsilon)$ is true at 0 and false at ε and $f(\varepsilon) = 0$, $\sim p(\varepsilon) \wedge \mathrm{B}p(\varepsilon)$ and $\sim p(\varepsilon) \wedge \sim \mathrm{B}{\sim} p(\varepsilon)$ are true at ε; since $f(2\varepsilon) = \varepsilon$, $\mathrm{B}(\sim p(\varepsilon) \wedge \mathrm{B}p(\varepsilon))$ and

$B(\sim p(\varepsilon) \wedge \sim B\sim p(\varepsilon))$ are true at 2ε. Thus at 2ε one has presumably irrational beliefs tantamount to '$p(\varepsilon)$ is false but I believe $p(\varepsilon)$' and '$\sim p(\varepsilon)$ is true but I don't believe $\sim p(\varepsilon)$'. There are several ways around this point. Most simply, we can make Bp true at w if and only if p is true at both $f(w)$ and 0. One believes oneself to be in $f(w)$ or 0 and is not sure which. Since one still believes all and only truths at 0, neither $\sim p \wedge Bp$ nor $p \wedge \sim Bp$ is ever true at 0, so neither $B(\sim p \wedge Bp)$ nor $B(p \wedge \sim Bp)$ is ever true at w. Under this modification, one's beliefs are still consistent and deductively closed at every world but complete only at 0. W can show almost as before that one believes no contingent proposition sensitively at w if $|w| \leq \varepsilon/2$.

We could also 'fill in' the worlds between $f(w)$ and 0 by making Bp true at w if and only if p is true at every world x such that $f(w) \leq x \leq 0$ or $0 \leq x \leq f(w)$. One believes oneself to be in some world between $f(w)$ and 0 inclusive and is not sure which. Under this modification too, one's beliefs are consistent and deductively closed at every world but complete only at 0. Moorean paradoxes are avoided and we can show in a similar way that one believes no contingent proposition sensitively at w if $|w| \leq \varepsilon/2$.

The last model differs from the first two in having a transitive accessibility relation and therefore validating all instances of the 'positive introspection' (S4) principle $Bp \supset BBp$; if one believes something then one believes that one believes it. In the previous two models, by contrast, since $p(2\varepsilon) \supset p(\varepsilon)$ is true at 2ε and 0 but not at ε, $B(p(2\varepsilon) \supset p(\varepsilon))$ is true at 3ε but not at 2ε, so $BB(p(2\varepsilon) \supset p(\varepsilon))$ is false at 3ε, so positive introspection fails at 3ε. If positive introspection holds everywhere in a model in which one's beliefs are everywhere consistent, wherever Bp is true so is BBp, so $\sim B\sim Bp$ is true by consistency, so $Bp \mathrel{\Box\!\!\rightarrow} \sim B\sim Bp$ is true everywhere: one cannot believe insensitively that one does not believe p. This is consistent with one's failure to believe contingent propositions sensitively, for in both the second and third models one believes $\sim Bp$ at some world only if one believes it at every world; if $B\sim Bp$ is true at w then $\sim Bp$ is true at 0, so p is false at 0, so $\sim Bp$ is true at every world; belief in it is vacuously sensitive.

None of the three models validates all instances of the 'negative introspection' (S5) principle that if one does not believe something then one believes that one does not believe it. For example, since $p(\varepsilon)$ is false at ε, $Bp(\varepsilon)$ is false at 2ε; since $p(\varepsilon)$ is true at 0, $Bp(\varepsilon)$ is true at ε, so $B\sim Bp(\varepsilon)$ is false at 2ε; thus $\sim Bp(\varepsilon) \supset B\sim Bp(\varepsilon)$ is false at 2ε. If negative introspection holds everywhere in a model in which one's beliefs are everywhere consistent, then wherever $\sim Bp$ is true so is $B\sim Bp$, so $\sim BBp$ is true by consistency, so $\sim Bp \mathrel{\Box\!\!\rightarrow} \sim BBp$ holds everywhere: one cannot believe insensitively that one believes p. In the models above, one can falsely believe that one believes p; $\sim Bp(\varepsilon) \wedge BBp(\varepsilon)$ is true at 2ε.

That the models are not realistic is obvious. Nevertheless, they enable us to see how sensitivity works, and the startling demands it can impose. They further undermine (4) in section 7.5, by showing how it can make the slightest bias destroy all knowledge of contingencies whatsoever. The same pathology can be expected in models of more realistic complexity, although of course harder to establish there. Intuitively, what goes wrong is that the counterfactual supposition ~*p* can take one to worlds at which one believes *p* on too different a basis from that on which one actually believes *p*. The obvious remedy is to relativize sensitivity to finely individuated methods, perhaps as (3) does in section 7.4.

APPENDIX 4
Iterated Probabilities in Epistemic Logic (Proofs)

First some standard definitions. A *frame* is a pair <W,R>, where R is a binary relation on the set W. For $w \in W$, $R(w) = \{x \in W: wRx\}$. R is *serial* on W just in case for every $w \in W$ there is an $x \in W$ such that wRx. <W,R> is serial (reflexive, symmetric, transitive) just in case R is serial (reflexive, symmetric, transitive) on W. A frame is *partitional* just in case it is reflexive, symmetric, and transitive. A probability distribution on W is a mapping P from all subsets of W to non-negative real numbers such that $P(W) = 1$ and $P(X \cup Y) = P(X) + P(Y)$ whenever X and Y are disjoint subsets of W. If $P(Y) > 0$, $P(X|Y) = P(X \cap Y)/P(Y)$; if $P(Y) = 0$, $P(X|Y)$ is undefined. A probability distribution P on W is *regular* just in case $P(X) > 0$ whenever X is a non-empty subset of W. A frame <W,R> is *bland* just in case it is serial and $P(X) = \sum_{w \in W} P(\{w\})P(X|R(w))$ for every $X \subseteq W$ and regular probability distribution P on W. $R(w)$ is always non-empty when <W,R> is serial, so $P(R(w)) > 0$ if P is regular, so $P(X|R(w))$ is defined; if <W,R> is not serial, $R(w)$ is empty for some $w \in W$ and $P(X|R(w))$ is undefined.

Proposition 1. Every bland finite frame is reflexive.

Proof. Let <W,R> be a bland finite frame and $x \in W$. Suppose that not xRx. Since <W,R> is serial, xRy for some $y \neq x$. Thus W has $n+2$ members for some $n \geq 0$. There is a (unique) regular probability distribution P on W such that:

$$P(\{x\}) = 2/3$$
$$P(\{w\}) = 1/3(n+1) \text{ for } w \in W - \{x\}.$$

Now

$$\sum_{w \in W} P(\{w\})P(W - \{x\}|R(w)) \geq P(\{x\})P(W - \{x\}|R(x)).$$

Since not xRx, $R(x) \subseteq W - \{x\}$, so $P(W - \{x\}|R(x)) = 1$. Thus

$$\sum_{w \in W} P(\{w\})P(W - \{x\}|R(w)) \geq P(\{x\}) = 2/3 > 1/3 = P(W - \{x\}).$$

This contradicts the blandness of <W,R>. Thus xRx. ■

Proposition 2. Every bland finite frame is symmetric.

Proof. Let <W,R> be a bland finite frame, and $x,y \in W$. Suppose that xRy but not yRx. Hence $x \neq y$, so W has $n+2$ members for some $n \geq 0$. There is a (unique) regular probability distribution P on W such that:

$$\begin{aligned}
&P(\{x\}) = 1/2\\
&P(\{y\}) = (n+2)/4(n+1)\\
&P(\{w\}) = 1/4(n+1) \text{ for } w \in W-\{x,y\}.
\end{aligned}$$

Now

$$\begin{aligned}
\textstyle\sum_{w\in W} P(\{w\})P(\{y\}|R(w)) &\geq P(\{x\})P(\{y\}|R(x)) + P(\{y\})P(\{y\}|R(y))\\
&= P(\{y\}|R(x))/2 + (n+2)P(\{y\}|R(y))/4(n+1)\\
&> P(\{y\}|R(x))/2 + P(\{y\}|R(y))/4.
\end{aligned}$$

Since xRy, $\{y\}\subseteq R(x)$, so

$$P(\{y\}|R(x)) = P(\{y\})/P(R(x)) \geq P(\{y\}).$$

But not yRx, so $R(y)\subseteq W-\{x\}$, so

$$P(R(y)) \leq P(W-\{x\}) = 1 - P(\{x\}) = 1/2.$$

Now yRy because R is reflexive by Proposition 1, so $\{y\}\subseteq R(y)$, so

$$P(\{y\}|R(y)) = P(\{y\})/P(R(y)) \geq 2P(\{y\}).$$

Thus

$$\textstyle\sum_{w\in W} P(\{w\})P(\{y\}|R(w)) > P(\{y\})/2 + 2P(\{y\})/4 = P(\{y\}).$$

This contradicts the blandness of <W,R>. Thus if xRy then yRx. ■

Proposition 3. Every bland finite frame is transitive.

Proof. Let <W,R> be a bland finite frame, and $x,y,z\in W$. Suppose that xRy and yRz but not xRz. Hence $x\neq y$ and $y\neq z$. By Proposition 1, xRx, so $x\neq z$. Thus W has $n+3$ members for some $n\geq 0$. There is a (unique) regular probability distribution P on W such that:

$$\begin{aligned}
&P(\{x\}) = P(\{z\}) = 1/3\\
&P(\{w\}) = 1/3(n+1) \text{ for } w\in W-\{x,z\}.
\end{aligned}$$

Now

$$\begin{aligned}
\textstyle\sum_{w\in W} P(\{w\})P(\{y\}|R(w)) &\geq P(\{x\})P(\{y\}|R(x)) + P(\{y\})P(\{y\}|R(y)) +\\
&\quad P(\{z\})P(\{y\}|R(z))\\
&= P(\{y\}|R(x))/3 + P(\{y\}|R(y))/3(n+1) + P(\{y\}|R(z))/3.
\end{aligned}$$

Since xRy, $\{y\}\subseteq R(x)$, so

$$P(\{y\}|R(x)) = P(\{y\})/P(R(x)).$$

Moreover, yRz and R is symmetric by Proposition 2, so zRy, so $\{y\}\subseteq R(z)$, so

$$P(\{y\}|R(z)) = P(\{y\})/P(R(z)).$$

Finally, yRy since R is reflexive, so $\{y\}\subseteq R(y)$, so

$$P(\{y\}|R(y)) = P(\{y\})/P(R(y)).$$

Thus

$$\begin{aligned}\textstyle\sum_{w\in W} P(\{w\})P(\{y\}|R(w)) &\geq P(\{y\})/3P(R(x)) + P(\{y\})/3(n+1)P(R(y)) + \\ &\quad P(\{y\})/3P(R(z)) \\ &> 1/3P(\{y\})(1/P(R(x)) + 1/P(R(z))).\end{aligned}$$

Since not xRz, $R(x)\subseteq W-\{z\}$, so

$$P(R(x)) \leq P(W-\{z\}) = 1 - P(\{z\}) = 2/3.$$

Similarly, not zRx because R is symmetric, so

$$P(R(z)) \leq 2/3.$$

Hence

$$1/3P(\{y\})(1/P(R(x)) + 1/P(R(z))) \geq 1/3P(\{y\})(3/2 + 3/2) = P(\{y\}).$$

Thus

$$\textstyle\sum_{w\in W} P(\{w\})P(\{y\}|R(w)) > P(\{y\}).$$

This contradicts the blandness of <W,R>, so if xRy and yRz then xRz. ■

Proposition 4. Every finite partitional frame is bland.

Proof. Let <W,R> be a finite partitional frame and P a regular probability distribution on W. Since R is reflexive on W, <W,R> is serial. Let $w,x\in W$. If not $w\in R(x)$ then not $x\in R(w)$ because R is symmetric, so $\{x\}\cap R(w) = \{\}$, so

$$P(\{x\}|R(w)) = 0.$$

If $w\in R(x)$ then $R(x) = R(w)$ and $\{x\}\subseteq R(w)$, since <W,R> is partitional, so

$$P(\{x\}|R(w)) = P(\{x\})/P(R(w)) = P(\{x\})/P(R(x)).$$

Hence

$$\begin{aligned}\textstyle\sum_{w\in W} P(\{w\})P(\{x\}|R(w)) &= \textstyle\sum_{w\in R(x)} P(\{w\})P(\{x\}|R(w)) \\ &= \textstyle\sum_{w\in R(x)} P(\{w\})P(\{x\})/P(R(x)) \\ &= \textstyle(P(\{x\})/P(R(x)))\sum_{w\in R(x)} P(\{w\}) \\ &= (P(\{x\})/P(R(x)))/P(R(x)) \\ &= P(\{x\}).\end{aligned}$$

Hence, for any $X\subseteq W$, $\sum_{w\in W} P(\{w\})P(X|R(w)) = P(X)$. ■

Proposition 5. A finite frame is bland if and only if it is partitional.

Proof. From Propositions 1-4. ■

Remark. The proofs of Propositions 1–3 and 5 use non-uniform distributions: $P(\{x\}) \neq P(\{y\})$ for some $x,y\in W$ (for finite frames, uniformity entails regularity). This is essential, in the sense that if 'bland' had been defined with 'uniform' in place of 'regular' then the analogues of Propositions 1–3 would have been false.

A non-partitional frame $<W,R>$ can satisfy the equation $P(X) = \sum_{w\in W} P(\{w\})P(X|R(w))$ for every $X\subseteq W$ when P is the uniform distribution on W. To see this, let $W = \{0, 1, 2\}$ and $R = \{<0,1>, <1,2>, <2,0>\}$. Thus $R(0) = \{1\}$, $R(1) = \{2\}$, $R(2) = \{0\}$, and R is serial but neither reflexive, symmetric, nor transitive on W. Let P be the uniform distribution on W; $P(\{0\}) = P(\{1\}) = P(\{2\}) = 1/3$. Nevertheless,

$$\begin{aligned} \sum_{w\in W} P(\{w\})P(X|R(w)) &= \sum_{w\in W} P(X|R(w))/3 \\ &= P(X\cap\{1\})/3P(\{1\}) + P(X\cap\{2\})/3P(\{2\}) + \\ &\quad P(X\cap\{0\})/3P(\{0\}) \\ &= P(X\cap\{1\}) + P(X\cap\{2\}) + P(X\cap\{0\}) \\ &= P(X). \end{aligned}$$

Of course, the examples in the main text show that not every finite serial frame is bland even in this weakened sense.

Now say that a frame $<W,R>$ is *banal* just in case it is serial and $P(X|\{w\in W:P(X|R(w))=c\}) = c$ for every real number c, subset X of W and regular probability distribution P on W such that the conditional probability is defined (i.e. $P(X|R(w))=c$ for some $w\in W$). Banality is a form of Miller's Principle or the Principal Principle.

Proposition 6. A finite frame is banal if and only if it is partitional.

Proof. Suppose that $<W,R>$ is a finite partitional frame, $X\subseteq W$ and P is a regular probability distribution on W. If wRx then $R(w) = R(x)$ since R is an equivalence relation, so $P(X|R(x))=c$ if $P(X|R(w))=c$. Thus, if the conditional probability is defined,

$$\{w\in W:P(X|R(w))=c\} = R(w_1) \cup \ldots \cup R(w_n)$$

where the $R(w_i)$ are pairwise disjoint and $P(X|R(w_i))=c$ for $1\leq i\leq n$. Thus

$$P(X\cap R(w_i)) = cP(R(w_i))$$

for $1\leq i\leq n$. Hence

$$\begin{aligned} P(X\cap\{w\in W:P(X|R(w))=c\}) &= P(X\cap(R(w_1) \cup \ldots \cup R(w_n)) \\ &= \textstyle\sum_{1\leq i\leq n} P(X\cap R(w_i)) \\ &= c\textstyle\sum_{1\leq i\leq n} P(R(w_i)) \\ &= cP(R(w_1) \cup \ldots \cup R(w_n)) \\ &= cP(\{w\in W:P(X|R(w))=c\}). \end{aligned}$$

Thus $P(X|\{w\in W:P(X|R(w))=c\}) = c$.

Conversely, suppose that $<W,R>$ is a finite banal frame. Let $W = \{w_0, \ldots, w_m\}$. There is a regular probability distribution P on W such that:

$$P(\{w_i\}) = 2^i/(2^{m+1}-1) \qquad (1\leq i\leq m)$$

Thus for all $X,Y\subseteq W$, $P(X) = P(Y)$ only if $X = Y$. Suppose that xRy. Let $c = P(\{y\}|R(x))$. Since $y\in R(x)$, $c>0$ and $P(\{y\}) = cP(R(x))$. Now suppose that $P(\{y\}|R(w)) = c$. Since $c>0$, $y\in R(w)$, so $P(\{y\}) = cP(R(w))$. Hence $cP(R(w)) =$

$cP(R(x))$; since $c>0$, $P(R(w)) = P(R(x))$, so $R(w)) = R(x)$. Conversely, if $R(w) = R(x)$ then $P(\{y\}|R(w)) = P(\{y\}|R(x)) = c$. Thus:

$$\{w{:}P(\{y\}|R(w))=c\} = \{w{:}R(w)=R(x)\}$$

Hence:

$$P(\{y\}|\{w{:}R(w)=R(x)\} = P(\{y\}|\{w{:}P(\{y\}|R(w))=c\}) = c = P(\{y\}|R(x))$$

because <W,R> is banal. By reasoning as above, $\{w{:}R(w)=R(x)\} = R(x)$. Since <W,R> is serial (because banal), this conclusion holds for all $x \in W$. Thus wRx if and only if $R(w)=R(x)$; since the latter equation defines an equivalence relation, <W,R> is partitional. ■

APPENDIX 5
A Non-Symmetric Epistemic Model

A creature stores information in sentential form. Its language L has two atomic sentences H ('It is hot') and C ('It is cold'), the logical constants $\sim$ and $\wedge$ with their usual interpretations, and the unary sentence functor K ('I know that ...'). Let <A> be the proposition expressed by the sentence A on this interpretation, and $W = \{w_1, w_2, x\}$. In order to specify which sentences are stored in which worlds, recursively define an auxiliary function ϕ from L to W:

$$\begin{array}{lll}
\phi H & = \{w_1\} & \\
\phi C & = \{w_2\} & \\
\phi{\sim}A & = W - \phi A & \\
\phi(A \wedge B) & = \phi A \cap \phi B & \\
\phi KA & = \phi A & \text{if } x \in \phi A \\
 & = \{\} & \text{otherwise.}
\end{array}$$

Let the creature be so connected to its environment that for all $y \in W$ and $A \in L$:

> # It is disposed in y to store A if and only if $y \in \phi KA$.

For example, since $\phi{\sim}C = \{w_1, x\}$, $\phi K{\sim}C = \{w_1, x\}$, so it is disposed to store $\sim$C in w_1 and x but not in w_2. Since $\phi KC = \{\}$, it is not disposed to store C in any world. Thus # agrees with the example in section 10.6 on the storage of information about whether it is cold; likewise for information about whether it is hot.

We can argue plausibly that ϕA is the set of worlds in which <A> is true. The argument is by induction on the complexity of A. The only non-routine case is the induction step for K. The induction hypothesis is that ϕA is the set of worlds in which <A> is true. Since the clause for ϕKA implies that $\phi KA \subseteq \phi A$, # implies that A is stored only in worlds in ϕA. Thus, by the induction hypothesis, the creature is disposed to store A only when <A> is true. We can therefore reasonably suppose that if the creature is disposed to store A then it knows <A>. Conversely, if it is not disposed to store A, then it does not know <A>. But <KA> is true if and only if it knows <A>. Thus <KA> is true if and only if the creature is disposed to store A. It follows by # that ϕKA is the set of worlds in which <KA> is true. This completes the induction step. ■

Given that ϕA is the set of worlds in which <A> is true, the definition of ϕ recursively specifies the truth-conditions of sentences of L. One can easily check that its results coincide with those of a semantics in possible worlds style, using the accessibility relation in the diagram in section 10.6. Since the accessibility relation is reflexive and transitive, every theorem of the modal system S4 is true in every world, when □ is replaced by K and propositional variables by arbitrary sentences of L. Consequently, the creature knows every logical consequence of what it knows; moreover, whenever it knows *p*, it knows that it knows *p*. But sometimes, when it does not know *p*, it does not know that it does not know *p*. That is because it cannot survey the totality of its knowledge. It is a failure of self-knowledge, not of rationality in any ordinary sense.

APPENDIX 6
Distribution over Conjunction

There are two obvious ways of trying to approximate an operator O which lacks a feature F by an operator with F. One is to seek the weakest operator O^+ stronger than O which has F; the other is to seek the strongest operator O^- weaker than O which has F. Of course, there is no general guarantee that either O^+ or O^- exists. However, when F is the feature of distributing over conjunction, we can define both O^+ and O^- in terms of O by quantifying into sentence position.

Formally, we use a binary sentence operator Con, where $\mathrm{Con}(p, q)$ is true if and only if p is semantically a conjunct of q. We understand this in such a way that all of the following are true:

$$\forall p \forall q \Box \mathrm{Con}(p, p \wedge q)$$

$$\forall p \forall q \Box \mathrm{Con}(q, p \wedge q)$$

$$\forall p \forall q \Box (\mathrm{Con}(p, q) \supset (q \supset p))$$

$$\forall p \Box \mathrm{Con}(p, p)$$

$$\forall p \forall q \forall r \Box ((\mathrm{Con}(p, q) \wedge \mathrm{Con}(q, r)) \supset \mathrm{Con}(p, r))$$

The operator O distributes over conjunction if and only if $\forall p \forall q \Box (\mathrm{Con}(p, q) \supset (\mathrm{O}q \supset \mathrm{O}p))$ is true. O is factive if and only if $\forall p \Box (\mathrm{O}p \supset p)$ is true. The operator O_1 entails the operator O_2 if and only if $\forall p \Box (\mathrm{O}_1 p \supset \mathrm{O}_2 p)$ is true. We assume that $\forall p$ commutes with $\Box$, i.e. that the Barcan formula and its converse hold; this is in effect to assume that it is not contingent what propositions there are.

We define the operators O^+ and O^- thus:

$$\mathrm{O}^+ p =_{\mathrm{def}} \forall q (\mathrm{Con}(q, p) \supset \mathrm{O}q)$$

$$\mathrm{O}^- p =_{\mathrm{def}} \exists q (\mathrm{Con}(p, q) \wedge \mathrm{O}q)$$

We first show that O^+ is the weakest operator at least as strong as O to distribute over conjunction. More precisely, we show: O^+ entails O; O^+ distributes over conjunction; if an operator O^* entails O and distributes over conjunction then O^* entails O^+. O^+ entails O because Con is reflexive. O^+ distributes over conjunction because, given $\mathrm{O}^+ p$ and $\mathrm{Con}(q, p)$, we can derive $\mathrm{O}^+ q$ since Con is transitive. Now suppose that O^* entails O and distributes over conjunction. We show that O^* entails O^+. Assume $\mathrm{O}^* p$. Given $\mathrm{Con}(q, p)$ we have $\mathrm{O}^* q$ because O^* distributes over conjunction, so we have $\mathrm{O}q$ because O^* entails O. Hence we have O^+p. Consequently, O^* entails O^+, as required. We also note that, since O^+ entails O, O^+ is factive if O is. ■

Given that K is factive, we can construct a Fitch-style argument from $\forall p(p \supset \Diamond K^+p)$ to $\forall p(p \supset K^+p)$. Since K^+ entails K, that yields an argument from $\forall p(p \supset \Diamond K^+p)$ to $\forall p(p \supset Kp)$. Since $\forall p(p \supset Kp)$ is absurd, we must deny $\forall p(p \supset \Diamond K^+p)$: a proposition can be true even if it is impossible to know all its conjuncts.

We now show that O^- is the strongest operator at least as weak as O to distribute over conjunction. More precisely, we show: O entails O^-; O^- distributes over conjunction; if O entails an operator O^* that distributes over conjunction then O^- entails O^*. O entails O^- because Con is reflexive. O^- distributes over conjunction because, given O^-p and Con(q, p), we can derive O^-q since Con is transitive. Now suppose that O entails O^* and O^* distributes over conjunction. We show that O^- entails O^*. Assume O^-p. Hence we can assume Con(p, q) and Oq for some q. We have O^*q because O entails O^*, so we have O^*p since O^* distributes over conjunction. Consequently, O^- entails O^*, as required. We also note that since conjunctions entail their conjuncts, O^- is factive if O is. ■

Given that K is factive, we can construct a Fitch-style argument from $\forall p(p \supset \Diamond K^-p)$ to $\forall p(p \supset K^-p)$. Since K entails K^-, that yields an argument from $\forall p(p \supset \Diamond Kp)$ to $\forall p(p \supset K^-p)$. The conclusion says that every truth is a conjunct of a known truth. Since that is false, we must deny $\forall p(p \supset \Diamond Kp)$.

Are K^+ and K^- well-defined? More specifically, does the quantification into sentence position create some kind of impredicativity, circularity, or paradox? We may suppose that a formula is interpreted by being assigned a subset of a set I as its semantic value. We can think of I as a set of possible worlds, indices, or the like, and the set assigned to a formula as the set of such items at which it is true. If a is an assignment of subsets of I to formulas, then $a(\forall p\Phi(p))$ is the intersection of $a^*(\Phi(p))$ for all such assignments a^* differing from a on sentence letters at most over p. Similarly, $a(\exists p\Phi(p))$ is the union of $a^*(\Phi(p))$ for all such assignments a^* differing from a on sentence letters at most over p. Such a semantics is formally unproblematic. The only danger is that it may yield a rather coarse-grained interpretation of Con and K. For example, we might say that a(Con(A, B)) is I if a(A) is the intersection of a(B) and some subset of I; otherwise a(Con(A, B)) is {}. But then a(Con(A, B)) is I if and only if a(A) is a subset of a(B); in effect, all consequences are treated as conjuncts. Thus to deny $\forall p(p \supset \Diamond K^+p)$ is in effect to say that a proposition can be true even if it is impossible to know all its consequences. The falsity of $\forall p(p \supset K^-p)$ is in effect the existence of a truth that is a consequence of no known truth. A finer-grained semantics might be desirable, but might raise problems for the semantics of the quantifiers. Nevertheless, the results under the coarse-grained semantics are already highly unfavourable to the verificationist conception.

BIBLIOGRAPHY

Achinstein, P. (ed.) 1983. *The Concept of Evidence*. Oxford: Oxford University Press.

Armstrong, D. M. 1973. *Belief, Truth and Knowledge*. Cambridge: Cambridge University Press.

Aumann, R. 1976. 'Agreeing to disagree.' *Annals of Statistics*, 4: 1236–9.

Austin, J. L. 1946. 'Other minds.' *Proceedings of the Aristotelian Society*, supp. 20: 148–87.

Austin, J. L. 1962. *Sense and Sensibilia*, ed. G. J. Warnock. Oxford: Oxford University Press.

Ayer, A. J. 1940. *The Foundations of Empirical Knowledge*. London: Macmillan.

Bacharach, M. O. L. 1985. 'Some extensions to a claim of Aumann in an axiomatic model of knowledge'. *Journal of Economic Theory*, 37: 167–90.

—— 1992a. 'The acquisition of common knowledge.' In Bicchieri and Dalla Chiara 1992.

—— 1992b. 'Backward induction and beliefs about oneself.' *Synthese*, 91: 247–84.

Barwise, J., and Perry, J. 1983. *Situations and Attitudes*. Cambridge, Mass.: MIT Press.

Basu, K. 1996. 'A paradox of knowledge and some related observations.' Unpublished typescript.

Bicchieri, C. 1989. 'Self-refuting theories of strategic interaction: a paradox of common knowledge.' *Erkenntnis*, 30: 69–85.

—— 1992. 'Knowledge-dependent games: backward induction.' In Bicchieri and Dalla Chiara 1992.

—— and Dalla Chiara, C. (eds.) 1992. *Knowledge, Belief and Strategic Interaction*. Cambridge: Cambridge University Press.

Black, M. 1952. 'Saying and disbelieving.' *Analysis*, 13: 25–33.

Boghossian, P. 1994. 'The transparency of mental content.' *Philosophical Perspectives*, 8: 33–50.

BonJour, L. 1985. *The Structure of Empirical Knowledge*. Cambridge, Mass.: Harvard University Press.

Bovens, L. 1997. 'The backward induction argument for the finite iterated prisoner's dilemmas and the surprise exam paradox.' *Analysis*, 57: 179–86.

Brandom, R. B. 1983. 'Asserting.' *Noûs*, 17: 637–50.

—— 1994. *Making it Explicit*. Cambridge, Mass.: Harvard University Press.

Burge, T. 1978. 'Buridan and epistemic paradox.' *Philosophical Studies*, 34: 21–35.

Burge, T. 1979. 'Individualism and the mental.' *Midwest Studies in Philosophy*, 4: 73–121.

—— 1984. 'Epistemic paradox.' *Journal of Philosophy*, 81: 5–29.

—— 1986a. 'Cartesian error and the objectivity of perception.' In Pettit and McDowell 1986.

—— 1986b. 'Individualism and psychology.' *Philosophical Review*, 95: 3–45.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Castell, P. 1996. 'Epistemic probability II.' *Proceedings of the Aristotelian Society*, supp. 70: 79–94.

Chellas, B. F. 1980. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.

Child, T. W. 1994. *Causality, Interpretation and the Mind*. Oxford: Clarendon Press.

Christensen, D. 1991. 'Clever bookies and coherent beliefs.' *Philosophical Review*, 100: 229–47.

—— 1992. 'Confirmational holism and Bayesian epistemology.' *Philosophy of Science*, 59: 540–57.

—— 1996. 'Dutch-book arguments depragmatized: epistemic consistency for partial believers.' *Journal of Philosophy*, 93: 450–79.

Cohen, S. 1988. 'How to be a fallibilist.' *Philosophical Perspectives*, 2: 91–123.

Conee, E., and Feldman, R. 1998. 'The generality problem for reliabilism.' *Philosophical Studies*, 89: 1–29.

Cozzo, C. 1994. 'What can we learn from the paradox of knowability?' *Topoi*, 13: 71–8.

Craig, E. J. 1990a. *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Clarendon Press.

—— 1990b. 'Three new leaves to turn over.' *Proceedings of the British Academy*, 76: 265–81.

Dancy, J. (ed.) 1988. *Perceptual Knowledge*. Oxford: Oxford University Press.

—— 1995. 'Arguments from illusion.' *Philosophical Quarterly*, 45: 421–38.

Daniels, C. B. 1988. 'Privacy and verification.' *Analysis*, 48: 100–2.

Davidson, D. 1986. 'A coherence theory of truth and knowledge.' In E. LePore (ed.), *Truth and Interpretation*. Oxford: Blackwell.

DeRose, K. 1991. 'Epistemic possibilities.' *Philosophical Review*, 100: 581–605.

—— 1995. 'Solving the skeptical problem.' *Philosophical Review*, 104: 1–52.

—— 1996. 'Knowledge, assertion and lotteries.' *Australasian Journal of Philosophy*, 74: 568–80.

Dewey, J. 1938. *Logic: The Theory of Inquiry*. New York: Henry Holt.

Diaconis, P., and Zabell, S. 1982. 'Updating subjective probability.' *Journal of the American Statistical Association*, 77: 822–30.

Dretske, F. I. 1970. 'Epistemic operators.' *Journal of Philosophy*, 67: 1007–23.

—— 1981a. *Knowledge and the Flow of Information*. Oxford: Blackwell.

—— 1981b. 'The pragmatic dimension of knowledge.' *Philosophical Studies*, 40: 363–78.

Dudman, V. 1992. 'Probability and assertion.' *Analysis*, 52: 204–11.

Dummett, M. A. E. 1977. *Elements of Intuitionism*. Oxford: Clarendon Press.

—— 1978. *Truth and Other Enigmas*. London: Duckworth.

—— 1981. *Frege: Philosophy of Language*, 2nd edn. London: Duckworth.

—— 1991. *The Logical Basis of Metaphysics*. London: Duckworth.

—— 1993. *The Seas of Language*. Oxford: Clarendon Press.

Earman, J. 1992. *Bayes or Bust?* Cambridge, Mass.: MIT Press.

—— 1993. 'Underdetermination, realism and reason.' *Midwest Studies in Philosophy*, 18: 19–38.

Edgington, D. 1985. 'The paradox of knowability.' *Mind*, 94: 557–68.

—— 1995. 'On conditionals.' *Mind*, 104: 235–329.

Evans, M. G. J. 1982. *The Varieties of Reference*, ed. J. H. McDowell. Oxford: Clarendon Press.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. 1995. *Reasoning about Knowledge*. Cambridge, Mass.: MIT Press.

Field, H. H. 1978. 'A note on Jeffrey conditionalization.' *Philosophy of Science*, 45: 361–7.

Fine, G. 1992. 'Inquiry in the *Meno*.' In R. Kraut (ed.), *The Cambridge Companion to Plato*. Cambridge: Cambridge University Press.

Fine, K. 1970. 'Propositional quantifiers in modal logic.' *Theoria*, 36: 336–46.

Fitch, F. B. 1963. 'A logical analysis of some value concepts.' *Journal of Symbolic Logic*, 28: 135–42.

Fodor, J. A. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Mass.: MIT Press.

—— 1987. *Psychosemantics*. Cambridge, Mass.: MIT Press.

—— 1994. *The Elm and the Expert: Mentalese and its Semantics*. Cambridge, Mass.: MIT Press.

—— 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fricker, E. M. 1999. 'Knowing is not a state of mind.' Unpublished typescript.

Fudenberg, D., and Tirole, J. 1991. *Game Theory*. Cambridge, Mass.: MIT Press.

Fumerton, R. 2000. 'Williamson on skepticism and evidence.' *Philosophy and Phenomenological Research*, 60: 629–35.

Gaifman, H. 1988. 'A theory of higher order probabilities.' In B. Skyrms and W. Harper (eds.), *Causation, Chance, and Credence*. Dordrecht: Kluwer.

Garber, D. 1980. 'Field and Jeffrey Conditionalization.' *Philosophy of Science*, 47: 142–5.

Gazdar, G. 1979. *Pramatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Geach, P. T. 1972. *Logic Matters*. Oxford: Blackwell.

Geanakoplos, J. 1989. 'Game theory without partitions, and applications to speculation and consensus.' Cowles Foundation Discussion Paper 914, Yale University.

—— 1992. 'Common knowledge, Bayesean learning and market speculation with bounded rationality.' *Journal of Economic Perspectives*, 6.4: 58–82.

Geanakoplos, J. 1994. 'Common knowledge.' In R. Aumann and S. Hart (eds.), *Handbook of Game Theory*, vol. 2. Leiden: Elsevier.

Gettier, E. 1963. 'Is justified true belief knowledge?' *Analysis*, 23: 121–3.

Gibbons, J. 1998. 'Truth in action.' Unpublished typescript.

Ginet, C. 1979. 'Performativitiy.' *Linguistics and Philosophy*, 3: 245–65.

Glymour, C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Goldman, A. 1976. 'Discrimination and perceptual knowledge.' *Journal of Philosophy*, 73: 771–91.

Goldstein, M. 1983. 'The prevision of a prevision'. *Journal of the American Statistical Association* 78: 817–19.

Green, M., and Hitchcock, C. 1994. 'Reflections on Reflection: Van Fraassen on belief.' *Synthese*, 98: 297–324.

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge, Mass.: Harvard University Press.

Guttenplan, S. 1994. 'Belief, knowledge and the origins of content.' *Dialectica*, 48: 287–305.

Hall, N. 1999. 'How to set a surprise exam.' *Mind*, 108: 647–703.

Hambourger, R. 1987. 'Justified assertion and the relativity of knowledge.' *Philosophical Studies*, 51: 241–69.

Harman, G. 1965. 'The inference to the best explanation.' *Philosophical Review*, 74: 88–95.

—— 1968. 'Knowledge, inference and explanation.' *American Philosophical Quarterly*, 5: 164–73.

—— 1973. *Thought*. Princeton: Princeton University Press.

—— 1980. 'Reasoning and evidence one does not possess.' *Midwest Studies in Philosophy*, 5: 163–82.

—— 1986. *Change in View: Principles of Reasoning*. Cambridge, Mass.: MIT Press.

Harrison, C. 1969. 'The Unanticipated Examination in view of Kripke's semantics for modal logic.' In J. W. Davis, D. J. Hockney, and W. K. Wilson (eds.), *Philosophical Logic*. Dordrecht: Reidel.

Hart, W. D. 1979. 'The epistemology of abstract objects: access and inference.' *Proceedings of the Aristotelian Society*, supp. 53: 152–65.

—— and McGinn, C. 1976. 'Knowledge and necessity.' *Journal of Philosophical Logic*, 5: 205–8.

Heal, B. J. 1974. 'Explicit performative utterances and statements.' *Philosophical Quarterly*, 24: 106–21.

Hedenius, I. 1963. 'Performatives.' *Theoria*, 29: 115–36.

Hempel, C. G. 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: The Free Press.

Hild, Matthias. 1997. 'Induction and the Dynamics of Belief.' D.Phil. thesis, Oxford University.

Hintikka, K. J. J. 1962. *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.

—— 1975. 'Impossible worlds vindicated.' *Journal of Philosophical Logic*, 4: 475–84.

Hinton, J. M. 1967. 'Visual experiences.' *Mind*, 76: 217–27.

—— 1973. *Experiences*. Oxford: Oxford University Press.

Howson, C. 1996. 'Epistemic probability I.' *Proceedings of the Aristotelian Society*, supp. 70: 63–77.

—— and Urbach, P. 1993. *Scientific Reasoning: The Bayesian Approach*, 2nd edn. Chicago: Open Court.

Hughes, G. E., and Cresswell, M. J. 1996. *A New Introduction to Modal Logic*. London: Routledge.

Humberstone, I. L. 1988. 'Some epistemic capacities.' *Dialectica*, 42: 183–200.

Hurley, S. L. 1998. *Consciousness in Action*. Cambridge, Mass.: Harvard University Press.

Hyman, J. 1999. 'How knowledge works.' *Philosophical Quarterly*, 49: 433–51.

Jackson, F. 1987. *Conditionals*. Oxford: Blackwell.

—— 1996. 'Mental causation.' *Mind*, 105: 377–413.

—— 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

—— and Pettit, P. 1995. 'Some content is narrow.' In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press.

Janaway, C. 1989. 'Knowing about surprises: a supposed antinomy revisited.' *Mind*, 98: 391–409.

Jeffrey, R. 1975. 'Carnap's empiricism.' in G. Maxwell and R. Anderson (eds.), *Induction, Probability, and Confirmation*. Minneapolis: University of Minnesota Press.

—— 1983. *The Logic of Decision*, 2nd ed. Chicago: University of Chicago Press.

Jones, O. R. 1991. 'Moore's Paradox, assertion and knowledge.' *Analysis*, 51: 183–6.

Kaplan, D. 1970. 'S5 with quantifiable propositional variables.' *Journal of Symbolic Logic*, 35: 355.

—— 1989. 'Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals.' In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. New York: Oxford University Press.

—— and Montague, R. 1960. 'A paradox regained.' *Notre Dame Journal of Formal Logic*, 1: 79–90.

Kaplan, M. 1985. 'It's not what you know that counts.' *Journal of Philosophy*, 82: 350–63.

Kim, J. 1993. *Supervenience and Mind*. Cambridge: Cambridge University Press.

Kitcher, P. 1983. *The Nature of Mathematical Knowledge*. Oxford: Oxford University Press.

Koons, R. C. 1992. *Paradoxes of Belief and Strategic Rationality*. Cambridge: Cambridge University Press.

Kvanvig, J. 1995. 'The knowability paradox and the prospects for anti-realism.' *Noûs*, 29: 481–500.

Kyburg, H., Jr. 1974. *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel.

Lackey, J. 1999. 'Testimonial knowledge and transmission.' *Philosophical Quarterly*, 49: 471–90.

Lemmon, E. J. 1962. 'On sentences verifiable by their use.' *Analysis*, 22: 86–9.

Lenzen, W. 1980. *Glauben, Wissen und Wahrscheinlichkeit*. Vienna: Springer.

Levi, I. 1967. 'Probability kinematics.' *British Journal for the Philosophy of Science*, 18: 197–209.

Lewis, D. K. 1970. 'General semantics.' *Synthese*, 22: 18–67. Page reference to reprinting in Lewis 1983.

—— 1973. *Counterfactuals*. Oxford: Blackwell.

—— 1975. 'Languages and language.' In K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science*, 7. Minneapolis: University of Minnesota Press. Page reference to reprinting in Lewis 1983.

—— 1979. 'Attitudes *de dicto* and *de se*.' *Philosophical Review*, 88: 513–43.

—— 1983. *Philosophical Papers*, vol. 1. Oxford: Oxford University Press.

—— 1996. 'Elusive knowledge.' *Australasian Journal of Philosophy*, 74: 549–67.

Lindström, S. 1997. 'Situations, truth and knowability: a situation-theoretic analysis of a paradox by Fitch.' In E. Ejerhed and S. Lindström (eds.), *Logic, Action and Cognition: Essays in Philosophical Logic*. Dordrecht: Kluwer.

Lipton, P. J. 1991. *Inference to the Best Explanation*. London: Routledge.

Loar, B. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.

Lowe, E. J. 1995. 'The truth about counterfactuals.' *Philosophical Quarterly*, 45: 41–59.

Luper-Foy, S. 1984. 'The epistemic predicament: knowledge, Nozickian tracking, and skepticism.' *Australasian Journal of Philosophy*, 62: 26–48.

—— 1987. *The Possibility of Knowledge: Nozick and his Critics*. Totowa, N.J.: Rowman & Littlefield.

McDowell, J. H. 1977. 'On the sense and reference of a proper name.' *Mind*, 86: 159–85.

—— 1980. 'Meaning, communication and knowledge.' In Z. van Straaten (ed.), *Philosophical Subjects: Essays Presented to P. F. Strawson*. Oxford: Clarendon Press.

—— 1982. 'Criteria, defeasibility, and knowledge.' *Proceedings of the British Academy*, 68: 455–479. Partly reprinted with revisions in Dancy 1988.

—— 1989. 'One strand in the private language argument'. *Grazer Philosophische Studien*, 33/34: 285–303.

—— 1994. *Mind and World*. Cambridge, Mass.: Harvard University Press.

—— 1995. 'Knowledge and the internal'. *Philosophy and Phenomenological Research*, 55: 877–93.

MacIntosh, J. J. 1984. 'Fitch's factives.' *Analysis*, 44: 153–8.

Mackie, J. L. 1980. 'Truth and knowability.' *Analysis*, 40: 90–3.

McLelland, J., and Chihara, C. 1975. 'The surprise examination paradox.' *Journal of Philosophical Logic*, 4: 71–89.

Maher, P. 1996. 'Subjective and objective confirmation.' *Philosophy of Science*, 63: 149–74.

Martin, M. G. F. 1997. 'The reality of appearances'. In R. M. Sainsbury (ed.), *Thought and Ontology*. Milan: FrancoAngeli.

Melia, J. 1991. 'Anti-realism untouched.' *Mind*, 100: 341–2.

Millar, A. 1991. *Reasons and Experience*. Oxford: Clarendon Press.

Milne, P. M. 1991. 'A dilemma for subjective Bayesians—and how to resolve it.' *Philosophical Studies*, 62: 307–14.

Moore, G. E. 1912. *Ethics*. London: Thornton Butterworth.

—— 1962. *Commonplace Book: 1919–1953*. London: Allen & Unwin.

Moser, P. K. 1989. *Knowledge and Evidence*. Cambridge: Cambridge University Press.

Mott, P. 1998. 'Margins for error and the sorites paradox.' *Philosophical Quarterly*, 48: 494–504.

Noonan, H. W. 1993. 'Object-dependent thoughts: a case of superficial necessity but deep contingency?' In J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press.

Nozick, R. 1981. *Philosophical Explanations*. Oxford: Oxford University Press.

Pagin, P. 1994. 'Knowledge of proofs.' *Topoi*, 13: 93–100.

Parzen, E. 1960. *Modern Probability Theory and its Applications*. New York: John Wiley & Sons.

Peacocke, C. A. B. 1981. 'Are vague predicates incoherent?' *Synthese*, 46: 121–41.

—— 1986. *Thoughts: An Essay on Content*. Oxford: Blackwell.

—— 1992. *A Study of Concepts*. Cambridge, Mass.: MIT Press.

—— 1993. 'Externalist explanation.' *Proceedings of the Aristotelian Society*, 93: 203–30.

—— 1999. *Being Known*. Oxford: Clarendon Press.

Peirce, C. S. 1932. *Collected Papers of Charles Sanders Peirce*, vol. 2: *Elements of Logic*, ed. C. Hartshorne and P. Weiss. Cambridge, Mass.: Harvard University Press.

Percival, P. 1990. 'Fitch and intuitionistic knowability.' *Analysis*, 50: 182–7.

—— 1991. 'Knowability, actuality and the metaphysics of context-dependence'. *Australasian Journal of Philosophy*, 69: 82–97.

Perner, J. 1993. *Understanding the Representational Mind*. Cambridge, Mass.: MIT Press.

Pettit, P. 1986. 'Broad-minded explanation and psychology.' In Pettit and McDowell 1986.

—— and McDowell, J. H. (eds.) 1986. *Subject, Thought and Context*. Oxford: Clarendon Press.

—— and Sugden, R. 1989. 'The backward induction paradox.' *Journal of Philosophy*, 86: 169–82.

Plantinga, A. 1982. 'How to be an anti-realist.' *Proceedings of the American Philosophical Association*, 56: 47–70.

—— 1993. *Warrant and Proper Function*. Oxford: Oxford University Press.

Prichard, H. A. 1950. *Knowledge and Perception*. Oxford: Clarendon Press.

Putnam, H. 1973. 'Meaning and reference.' *Journal of Philosophy*, 70: 699–711.

—— 1978. *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

—— 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1969. 'Epistemology naturalized.' In his *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Rabinowicz, W., and Segerberg, K. 1994. 'Actual truth, possible knowledge.' *Topoi*, 13: 101–15.

Radford, C. 1966. 'Knowledge—by examples.' *Analysis*, 27: 1–11.

Recanati, F. 1987. *Meaning and Force: The Pragmatics of Performative Utterances*. Cambridge: Cambridge University Press.

Rescher, N. 1968. *Topics in Philosophical Logic*. Dordrecht: Reidel.

—— 1984. *The Limits of Science*. Berkeley: University of California Press.

Routley, R. 1981. 'Necessary limits to knowledge: unknowable truths.' In E. Morscher, O. Neumaier, and G. Zecha (eds.), *Essays in Scientific Philosophy*. Bad Reichenhall: Comes.

Rubinstein, A. 1989. 'The electronic mail game: strategic behavior under "almost common knowledge".' *American Economic Review*, 79: 385–91.

—— 1998. *Modeling Bounded Rationality*. Cambridge, Mass.: MIT Press.

Russell, B. A. W. 1910–11. 'Knowledge by acquaintance and knowledge by description.' *Proceedings of the Aristotelian Society*, 11: 108–28. Page reference to reprinting in Salmon and Soames 1988.

—— 1993. *Our Knowledge of the External World*. London: Routledge. First published 1914.

Sainsbury, R. M. 1995. *Paradoxes*, 2nd ed. Cambridge: Cambridge University Press.

—— 1997. 'Easy possibilities.' *Philosophy and Phenomenological Research*, 57: 907–19.

Salmon, N. U. 1982. *Reference and Essence*. Oxford: Blackwell.

—— 1986. 'Modal paradox: parts and counterparts, points and counterpoints.' in P. A. French, T. E. Uehling, and H. K. Wettstein (eds.), *Midwest Studies in Philosophy*, 11: *Studies in Essentialism*. Minneapolis: University of Minnesota Press.

—— 1989. 'The logic of what might have been.' *Philosophical Review*, 98: 3–34.

—— 1993. 'This side of paradox.' *Philosophical Topics*, 21: 187–97.

—— and Soames, S. (eds.). 1988. *Propositions and Attitudes*. Oxford: Oxford University Press.

Samet, D. 1990. 'Ignoring ignorance and agreeing to disagree.' *Journal of Economic Theory*, 52: 190–207.

Schiffer, S. R. 1996. 'Contextualist solutions to scepticism.' *Proceedings of the Aristotelian Society*, 96: 317–33.

Schlesinger, G. N. 1985. *The Range of Epistemic Logic*. Aberdeen: Aberdeen University Press.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Shatz, D. 1987. 'Nozick's conception of skepticism.' In Luper-Foy 1987.

Shaw, R. 1958. 'The paradox of the unexpected examination.' *Mind*, 67: 382–4.

Shin, H. S. 1989. 'Non-partitional information on dynamic state spaces and the possibility of speculation.' Center for Research on Economic and Social Theory Working Paper 90–11, University of Michigan.

—— 1992. Review of B. Skyrms, *The Dynamics of Rational Deliberation. Economics and Philosophy*, 8: 176–83.

—— 1993. 'Logical structure of common knowledge.' *Journal of Economic Theory*, 60: 1–13.

—— and Williamson, T. 1994. 'Representing the knowledge of Turing Machines.' *Theory and Decision*, 37: 125–46.

—— and Williamson, T. 1996. 'How much common belief is necessary for a convention?' *Games and Economic Behavior*, 13: 252–68.

Shope, R. K. 1978. 'The conditional fallacy in modern philosophy.' *Journal of Philosophy*, 75: 397–413.

—— 1983. *The Analysis of Knowing: A Decade of Research*. Princeton: Princeton University Press.

Skyrms, B. 1980. 'Higher order degrees of belief.' In D. H. Mellor (ed.), *Prospects for Pragmatism*. Cambridge: Cambridge University Press.

—— 1983. 'Three ways to give a probability assignment a memory.' In J. Earman (ed.), *Testing Scientific Theories*. Minneapolis: University of Minnesota Press.

—— 1987. 'Dynamic coherence and probability kinematics.' *Philosophy of Science*, 54: 1–20.

—— 1993. 'A mistake in dynamic coherence arguments?' *Philosophy of Science*, 60: 320–8.

Slote, M. A. 1979. 'Assertion and belief.' In J. Dancy (ed.), *Papers on Language and Logic*. Keele: Keele University Library.

Smith, M. 1994. *The Moral Problem*. Oxford, Blackwell.

Smith, P. 1984. 'Could we be brains in a vat?' *Canadian Journal of Philosophy*, 14: 115–23.

Snowdon, P. 1980–1. 'Perception, vision and causation.' *Proceedings of the Aristotelian Society*, 81: 175–92.

—— 1990. 'The objects of perceptual experience.' *Proceedings of the Aristotelian Society*, supp. 64: 121–50.

Soames, S. 1998. 'The modal argument: wide scope and rigidified descriptions.' *Noûs*, 32: 1–22.

Sorensen, R. A. 1988. *Blindspots*. New York: Oxford University Press.

Sosa, E. 1996. 'Postscript to "Proper fuctionalism and virtue epistemology".' In J. L. Kvanvig (ed.), *Warrant in Contemporary Epistemology*. Lanham, Md.: Rowman & Littlefield.

—— 2000. 'Contextualism and skepticism.' In J. Tomberlin, ed., *Philosophical Issues*: supp. to *Noûs*, 34.

Stalnaker, R. C. 1984. *Inquiry*. Cambridge, Mass.: MIT Press.

—— 1999. *Context and Content*. Oxford: Oxford University Press.

Stampe, D. 1987. 'The authority of desire.' *Philosophical Review*, 96: 335–81.

Steiner, M. 1975. *Mathematical Knowledge*. Ithaca, N.Y.: Cornell University Press.

Steup, M. 1992. 'Memory.' In J. Dancy and E. Sosa (eds.), *A Companion to Epistemology*. Oxford: Blackwell.

Steward, H. 1997. *The Ontology of Mind: Events, Processes and States*. Oxford: Clarendon Press.

Stich, S. P. 1978. 'Autonomous psychology and the belief-desire thesis.' *The Monist*, 61: 573–91.

Stine, G. C. 1976. 'Skepticism, relevant alternatives and deductive closure.' *Philosophical Studies*, 29: 249–61.

Talbott, W. J. 1991. 'Two principles of Bayesian epistemology.' *Philosophical Studies*, 62: 135–50.

Tennant, N. 1997. *The Taming of the True*. Oxford: Clarendon Press.

Thijsse, E. G. C. Forthcoming. 'The doxastic-epistemic force of declarative utterances.' In W. J. Black and H. C. Bunt (eds.), *Abduction, Beliefs and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam: John Benjamins.

Unger, P. 1972. 'Propositional verbs and knowledge.' *Journal of Philosophy*, 69: 301–12.

—— 1975. *Ignorance: A Case for Scepticism*. Oxford: Oxford University Press.

Usberti, G. 1995. *Significato e Conoscenza: Per una Critica del Neoverificazionismo*. Milan: Guerini Scientifica.

Van Fraassen, B. 1984. 'Belief and the will.' *Journal of Philosophy*, 81: 235–56.

—— 1995. 'Belief and the problem of Ulysses and the sirens.' *Philosophical Studies*, 77: 7–37.

Vendler, Z. 1967. *Linguistics in Philosophy*. Ithaca, N.Y.: Cornell University Press.

—— 1972. *Res Cogitans*. Ithaca, N.Y.: Cornell University Press.

Vogel, J. 1987. 'Tracking, closure and inductive knowledge.' In Luper-Foy 1987.

Weinstein, S. 1983. 'The intended interpretation of intuitionistic logic.' *Journal of Philosophical Logic*, 12: 261–70.

Weintraub, R. 1995. 'The Surprise Examination paradox.' *Ratio*, 8: 161–9.

Williams, B. A. O. 1978. *Descartes: The Project of Pure Enquiry*. London: Penguin.

Williamson, T. 1982. 'Intuitionism disproved?' *Analysis*, 42: 203–7.

—— 1987a. 'On knowledge and the unknowable.' *Analysis*, 47: 154–8.

—— 1987b. 'On the paradox of knowability.' *Mind*, 96: 256–61.

—— 1988. 'Knowability and constructivism.' *Philosophical Quarterly*, 38: 422–32.

—— 1990a. *Identity and Discrimination*. Oxford: Blackwell.

—— 1990b. 'Two incomplete anti-realist modal epistemic logics.' *Journal of Symbolic Logic*, 55: 297–314.

—— 1992a. 'An alternative rule of disjunction in modal logic.' *Notre Dame Journal of Formal Logic*, 33: 89–100.

—— 1992b. 'Inexact knowledge.' *Mind*, 101: 217–42.

—— 1992c. 'On intuitionistic modal epistemic logic.' *Journal of Philosophical Logic*, 21: 63–89.

—— 1993. 'Verificationism and non-distributive knowledge.' *Australasian Journal of Philosophy*, 71: 78–86.

—— 1994a. 'Never say never.' *Topoi*, 13: 135–45.

—— 1994b. *Vagueness*. London: Routledge.

—— 1995a. 'Does assertibility satisfy the S4 axiom?' *Crítica*, 27: 3–22.

—— 1995b. 'Is knowing a state of mind?' *Mind*, 104: 533–65.

—— 1996a. 'Cognitive homelessness.' *Journal of Philosophy*, 93: 554–73.

—— 1996b. 'Knowing and asserting.' *Philosophical Review*, 105: 489–523.

—— 1997. 'Knowledge as evidence.' *Mind*, 106: 717–41.

—— 1998a. 'Bare possibilia'. *Erkenntnis*, 48: 257–73.

—— 1998b. 'The broadness of the mental: some logical considerations.' In J. Tomberlin (ed.), *Philosophical Perspectives*, 12: *Language, Mind, and Ontology*. Oxford: Blackwell.

—— 1998c. 'Conditionalizing on knowledge.' *British Journal for the Philosophy of Science*, 49: 89–121.

—— 1999. 'Truthmakers and the converse Barcan formula.' *Dialectica*, 53: 253–70.

—— 2000a. 'Existence and contingency.' *Proceedings of the Aristotelian Society*, 100: 117–39.

—— 2000b. 'Margins for error: a reply.' *Philosophical Quarterly*, 50: 76–81.

—— 2000c. 'Skepticism and evidence.' *Philosophy and Phenomenological Research*, 60: 613–28.

—— 2000d. 'Skepticism, semantic externalism and Keith's mom.' *Southern Journal of Philosophy*, 38.

—— 2000e. 'Tennant on knowability.' *Ratio*, 13: 99–114.

Wittgenstein, L. 1958. *Philosophical Investigations*, 2nd ed., ed. G. E. M. Anscombe and R. Rhees, trans. G. E. M. Anscombe. Oxford: Blackwell.

—— 1969. *On Certainty*, ed. G. E. M. Anscombe and G. H. von Wright, trans. D. Paul and G. E. M. Anscombe. Oxford: Blackwell.

Wright, C. J. G. 1983. 'Keeping track of Nozick.' *Analysis*, 43: 134–40.

—— 1991. 'Scepticism and dreaming: imploding the demon.' *Mind*, 100: 87–116.

—— 1992a. 'On Putnam's proof that we are not brains-in-a-vat.' *Proceedings of the Aristotelian Society*, 92: 67–94.

—— 1992b. *Truth and Objectivity*. Cambridge, Mass.: Harvard University Press.

—— 1993. *Realism, Meaning and Truth*, 2nd ed. Oxford: Blackwell.

—— 1996. 'Response to Commentators.' *Philosophy and Phenomenological Research*, 56: 911–41.

—— and Sudbury, A. 1977. 'The paradox of the unexpected examination.' *Australasian Journal of Philosophy*, 55: 41–58.

Wright, G. H. von. 1957. *Logical Studies*. London: Routledge.

Yablo, S. 1992. 'Mental causation.' *Philosophical Review*, 101: 245–80.

—— 1997. 'Wide causation.' In J. Tomberlin (ed.), *Philosophical Perspectives*, 11: *Mind, Causation, and World*. Oxford: Blackwell.

Zemach, E. 1987. 'Are there logical limits for science?' *British Journal for the Philosophy of Science*, 38: 527–32.

INDEX